

For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex libris
UNIVERSITATIS
ALBERTAE NSIS



6m-31



Digitized by the Internet Archive
in 2020 with funding from
University of Alberta Libraries

<https://archive.org/details/Gaza1980>

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Hartmut K. von Gaza
TITLE OF THESIS Minimum Cost Paths for Inhomogeneous
 Cost Surfaces
DEGREE FOR WHICH THESIS WAS PRESENTED Master of Science
YEAR THIS DEGREE GRANTED 1980

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

THE UNIVERSITY OF ALBERTA

Minimum Cost Paths for Inhomogeneous Cost Surfaces

by

Hartmut K. von Gaza



A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF Master of Science

Geography

EDMONTON, ALBERTA

1980

THE UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Minimum Cost Paths for Inhomogeneous Cost Surfaces submitted by Hartmut K. von Gaza in partial fulfilment of the requirements for the degree of Master of Science.



ABSTRACT

This thesis deals with the problem of finding the minimum cost path between two points on a plane, for which an inhomogeneous cost is defined. It is a problem in the calculus of variations and requires the minimization of a cost-distance integral. The solving of the resulting Euler-Lagrange equations is avoided by representing the integral with an objective function which can then be minimized using a variety of mathematical optimization techniques. The decision variables for this objective function define the location of the links of a discretized path, and therefore a minimum of the function provides the minimum cost path. This method can be used for different cost surfaces and for paths subjected to both linear and non-linear constraints.

The case studies presented consider a number of surfaces and constraints. The cost surface represented by $C(x,y)=K\exp(-Ax^2-By^2)$ is used as a theoretical model to demonstrate the interaction between cost and distance. Minimum time paths for the velocity surfaces of Wardrop (1969), and Angel and Hyman (1972b) are calculated. In addition, three cases deal with imposing linear and non-linear constraints on the path.

Current literature explores the use of transformation methods to derive minimum cost paths. Puu (1978b) considers transforming the cost surface into another surface which upon distance equals cost. The minimum distance path defined

as the geodesic is shown to represent the minimum cost path. The relationship between the geodesic and the minimum cost path is explored and discussion of the transformation method is provided.

It is generally concluded that solving path related problems by minimizing objective functions is realistic and feasible.

ACKNOWLEDGEMENTS

I would like to acknowledge all those people who have helped me in the preparation of this thesis. John Hodgson has supplied me with years of encouragement, criticism, and patience. Hans Hosli, John Archer, John Rankin, Jeff Butler, and especially Henry Wolkowicz provided their mathematical expertise. Dietrich Wittkowski and Jean-Claude Muller contributed much theoretical discussion. Al Schaeffer, who read my thesis, insisted that the best minimum paths are the rivers of the north. The people at Edmonton Transit helped prepare initial drafts and provided me with other resources. Finally, I would like to thank my wife, Margriet, for her understanding when progress was slow.

TABLE OF CONTENTS

CHAPTER		PAGE
Chapter 1	The Minimum Cost Path Problem	1
1.1	Introduction	1
1.2	The minimum cost path	3
Chapter 2	Known Solution Methods	12
2.1	Early mathematical development of path problems .	12
2.2	General mathematical formulation	14
2.2.1	Continuous path formulations	15
2.2.2	Discrete formulations	16
2.2.3	Alternative methods	19
2.3	Discussion	21
2.4	The mathematical optimization formulation	25
Chapter 3	The Objective Function: Implementation and soltuion	27
3.1	Mathematical optimization	27
3.2	Modelling the minimum cost path as an objective function	32
3.2.1	The two link path and the objective function.....	33
3.2.2	Variable x and y decision variables	38
3.2.3	The variable r and theta method	38
3.2.4	A comparison of the two methods	44
3.2.5	Numerical integration analysis	48
3.2.6	Other solution criteria	51

TABLE OF CONTENTS continued

3.3 Constrained solution paths	54
Chapter 4 Examples of Minimum Cost paths	56
4.1 Variations on one cost surface	56
4.1.1 Cost surface orientation	57
4.1.2 The effect of cost on path curvature	64
4.1.3 Varying the base cost	68
4.2 Radially symmetric cost surfaces	68
4.2.1 Wardrop's velocity surface	71
4.2.2 Angel and Hyman's velocity surface	72
4.3 Constrained minimum cost paths	75
4.3.1 The barrier problem	76
4.3.2 Minimum cost paths contained within a corridor	76
4.3.3 Fixed distance penalty functions	79
Chapter 5 Geodesics and Minimum Cost Paths	83
5.1 Calculating the geodesic	83
5.1.1 Definition and objective function	83
5.1.2 Geodesic examples	85
5.1.3 The geodesic and the minimum cost path compared	87
5.1.4 Geodesics as minimum cost paths	89
5.2 Transformations, geodesics and minimum cost paths.....	92
5.2.1 Puu's contribution	94
5.2.2 The three steps in the transformation process	96
5.2.3 Results of the transformation methods	99

TABLE OF CONTENTS continued

Chapter 6 Conclusion101

 6.1 The methodology101

 6.1.1 The method works101

 6.1.2 The constraints103

 6.2 The transformation method105

 6.3 Future research106

 6.3.1 The cost surface106

 6.3.2 The use of constraints106

 6.3.3 Other path problems107

Bibliography108

Appendix I114

LIST OF TABLES

Table	Description	Page
1	Comparison of paths with varying number of links	50
2	Cost and distance analysis for Figure 16	60
3	Cost and distance analysis for Figure 18	61
4	Cost and distance analysis for Figure 20	67
5	Cost and distance analysis for Figure 22	68

LIST OF FIGURES

Figure		Page
1	Two paths for a homogeneous cost surface	5
2	The profile of cost versus distance from one end point for the paths in Figure 1	6
3	Two paths for an inhomogeneous cost surface	7
4	Profile of cost versus distance from one end point for paths in Figure 3	8
5	Contours and two link path for objective function (18)	35
6	Contours and two link path for objective function (20)	36
7	A path calculated using the x,y method	39
8	A five link path with equal but variable link length	40
9	A path calculated using the r and theta method	43
10	Paths calculated using the x,y and r, theta methods	45
11	Cost calculated at different number of links	46
12	Number of iterations required for convergence at different number of links	47
13	Paths calculated with 2, 4, 8, and 16 links	49
14	Two initial solutions - one optimum solution	52
15	Two symmetrical solutions	53
16	Four paths traversing the cost surface	58
17	Profile of cost versus distance from one end point for paths in Figure 16	59
18	Four paths avoiding the high cost regions	62
19	Profile of cost versus distance from one of the endpoint for paths in Figure 18	63

LIST OF FIGURES continued

20	The effect of cost on path curvature	65
21	Profile of cost versus distance from one end point for paths in Figure 20	66
22	The increasing base cost (BC) example	69
23	Profile of cost versus distance from one end point for paths in Figure 22	70
24	Minimum time paths for Wardrop's velocity surface	73
25	Minimum time paths for Angel and Hyman's velocity surface	74
26	Path constrained by a barrier	77
27	Two paths contained within a corridor	78
28	Paths constrained by circular regions	80
29	The Edmonton - London geodesic	86
30	Geodesic on a cone	88
31	Geodesic and the minimum cost path compared	90
32	Geodesic and the minimum cost path compared	91
33	Three geodesics on a sinuous surface	93

Chapter 1 The Minimum Cost Path Problem

1.1 Introduction

Locating optimal paths in geographic space is a problem often encountered by man and much energy is directed towards its solution. Nearly every person undertaking travel seeks to optimize some factor such as travel time, distance, cost, or even available scenery. Attempts at path optimization are evident in the type and location of transportation facilities man has constructed. From the short cuts seen in student travel patterns on university campuses, to interconnected modern airline routes, man attempts to find more efficient transportation routes. With increasing energy costs of providing both facilities and transport, the optimal location of transportation routes is of increasing interest to the geographer, planner, and engineer.

Optimal paths have been observed in the movement of nearly all natural phenomena such as the passage of light through various media, the meandering of streams, animal paths of least energy expenditure, and avalanche paths on mountain slopes. In the past the theoretical analysis of these paths has been left more to the physicist and mathematician than to the geographer. The mathematical analysis required to explain the location of optimal paths has been beyond the interest and capabilities of most geographers. Current geography, however, requires the methods of other disciplines to advance its own theoretical

base. The main body of theory concerning minimum cost paths occurs in physics and mathematics. As shown by Angel and Hyman (1976) and more recently by Puu (1978 a,b,c), the application of optimal path theory is of increasing concern to geographers.

This study is concerned with the methodology of finding optimal paths connecting two points on a plane. Optimal, in this sense, is defined in terms of minimizing some function of transportation cost. The design of this thesis is as follows. Chapter 1 describes the minimum path problem. Chapter 2 develops the theoretical framework for finding minimum cost paths and provides a brief review of the literature. Chapter 3 explains the techniques of mathematical optimization used to find the minimum cost path. Examples of both unconstrained and constrained minimum cost paths are presented in Chapter 4. The relationship between geodesics and minimum cost paths is examined in Chapter 5. The concluding Chapter 6 provides a discussion on the further development of minimum cost path methodologies.

1.2 The minimum cost path

The optimal path problem concerns locating a path of minimum cost connecting two points on a plane. Before such a path can be calculated two criteria must be met. First, the path must be free to locate anywhere on the Euclidean plane representing a general physical surface. Secondly, a cost surface which provides a transportation cost for every point on the plane must be defined. Many types of transportation cost variables have been modelled and these can represent the cost surface. Angel and Hyman (1976) use travel velocity as the inverse of a cost surface variable, Turner (1971) uses a combination of construction and land acquisition cost variables, and Warntz (1965) uses population potential as a land acquisition cost variable. Conceivably, a combination of travel, construction and time costs, such as Wilson's (1974) "generalized cost function", could be incorporated into the cost surface. This research does not concentrate on the development of cost surfaces; it assumes them to be given. Rather, it pursues the problem of finding minimum cost paths for such surfaces.

Once a cost surface and a path have been defined, the total cost of the path can be calculated. For the homogeneous or constant valued cost surface the total cost is calculated by multiplying the distance of the path by the cost per unit distance. The cost of a path for the inhomogeneous or non constant valued cost surface is calculated similarly. Rather than multiplying the total

distance of the path by one cost, each part of the path must be multiplied by the cost attributed to that part and summed for the total. This calculation is equivalent to finding the perpendicular cross-sectional area above the path which extends between the plane and the cost surface.

On a geographic plane for which a homogeneous cost surface is defined the minimum cost paths are straight lines. Figure 1 demonstrates two paths for a homogeneous cost surface. The constant value of this surface is given by the formulation $C(x,y)=K$, where x and y are coordinates and K is a constant equal to 5.0. The parameters "A", "B", and "BC" which are all equal to 0.0 are used to formulate more complex cost surfaces discussed later. Two paths are demonstrated in this figure. Path 1 consists of two links with a total length "DIST" of 7.80 and a total cost "COST" of 39.02. Path 2 connects the two points with a straight line of length 5.66 and cost of 28.28. Each of these costs can be derived by multiplying the distance by the constant value of the cost surface. Figure 2 illustrates the area or total cost of each path. Path 2, the straight line, is the path with the least area and cost.

Minimum cost paths for inhomogeneous cost surfaces are usually not straight. Figure 3 shows two paths on an inhomogeneous cost surface. The surface is projected onto the plane by an isomap in which isocost lines join points of equal cost per unit distance. In addition to the variable cost surface each path is subjected to a constant base cost

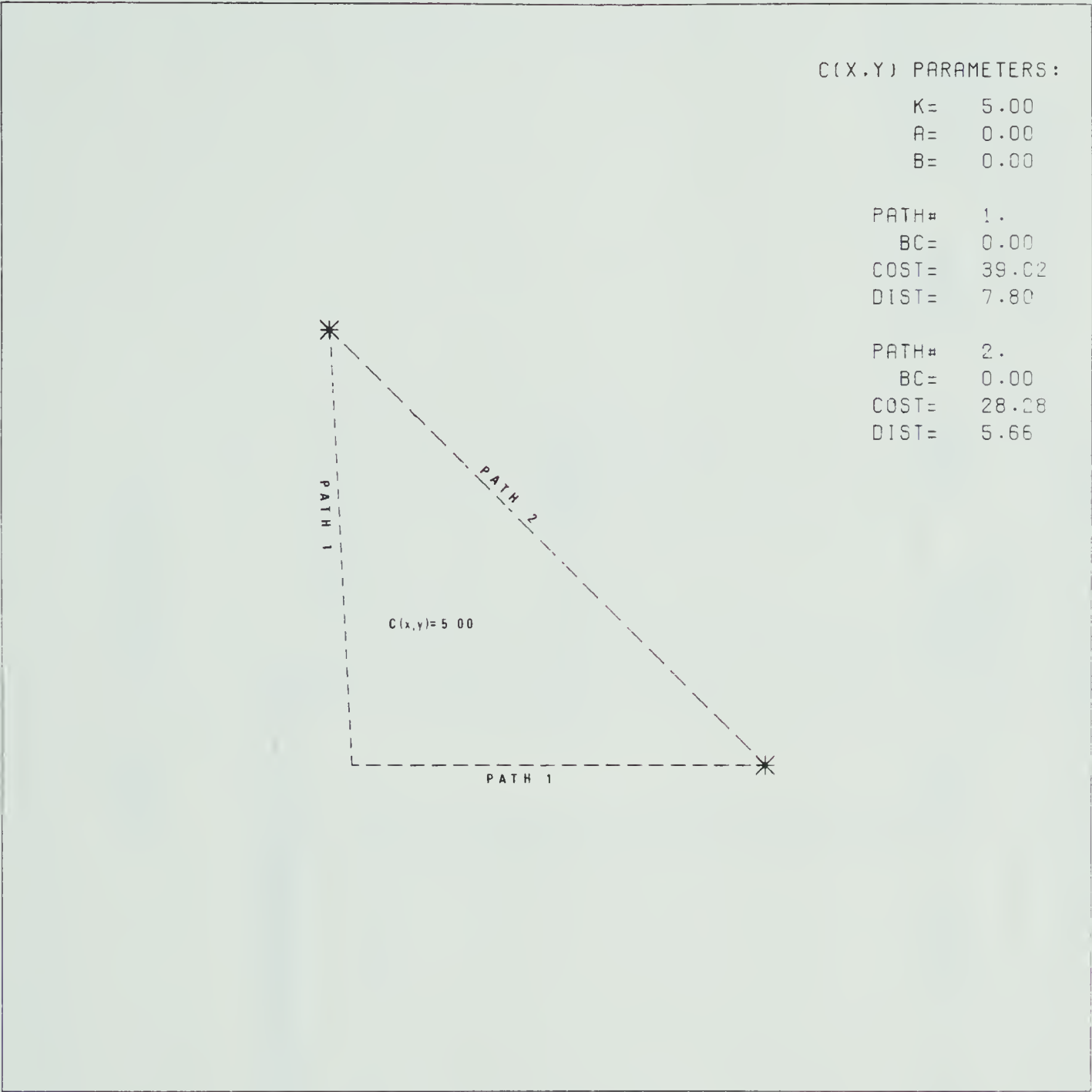


Figure 1 Two paths for a homogeneous cost surface

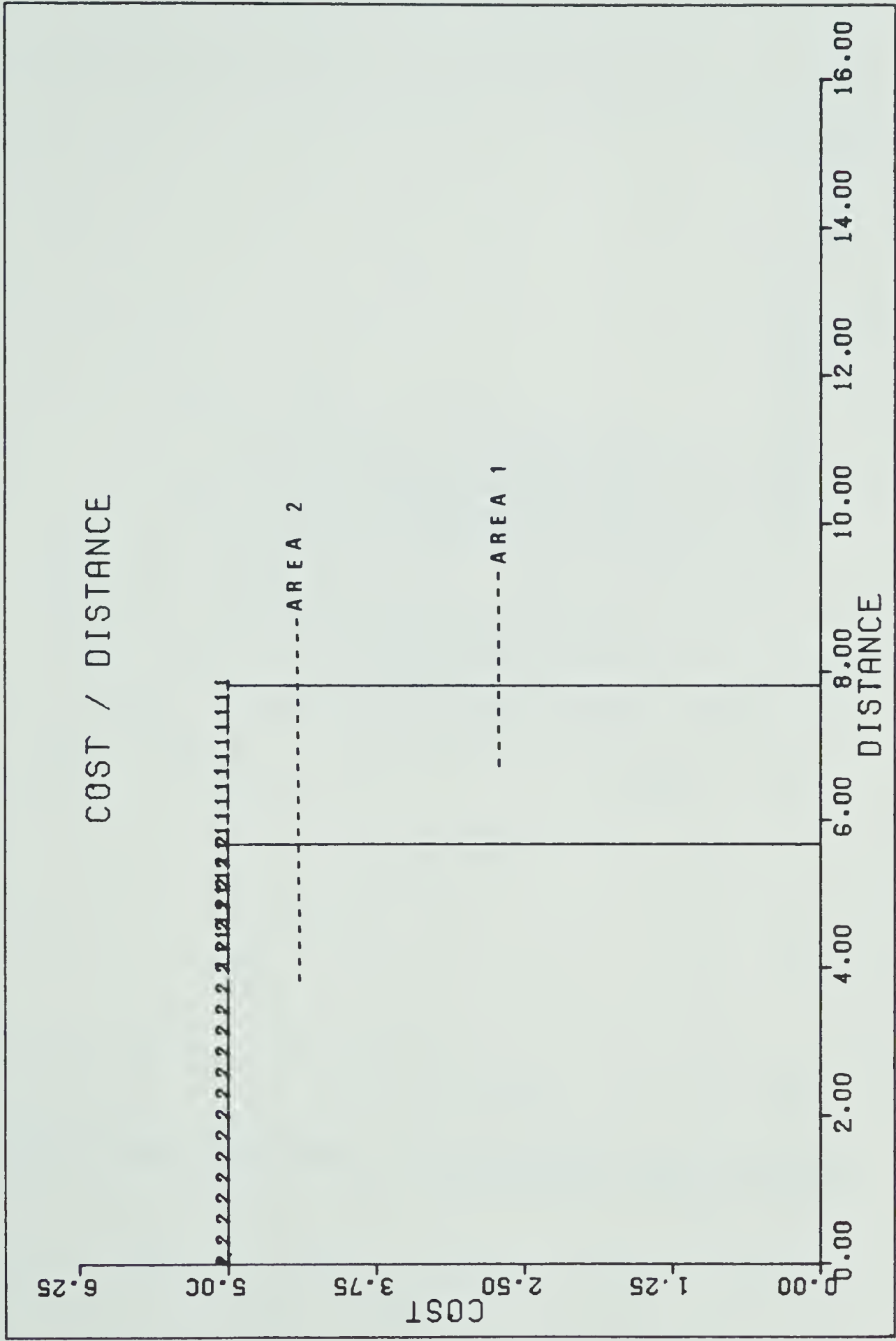


Figure 2 The profile of cost versus distance from one end point for the paths in Figure 1

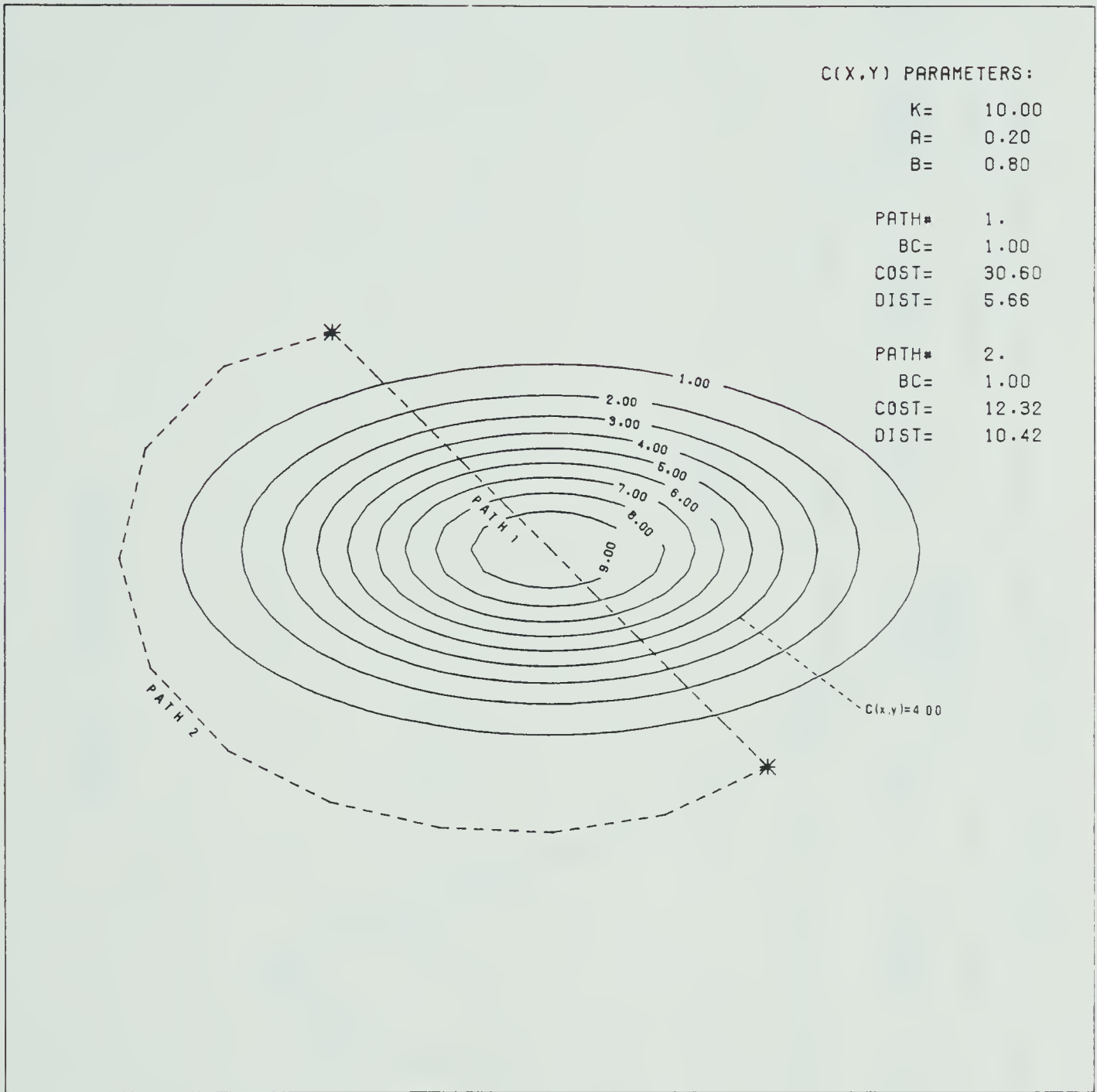


Figure 3 Two paths for an inhomogeneous cost surface

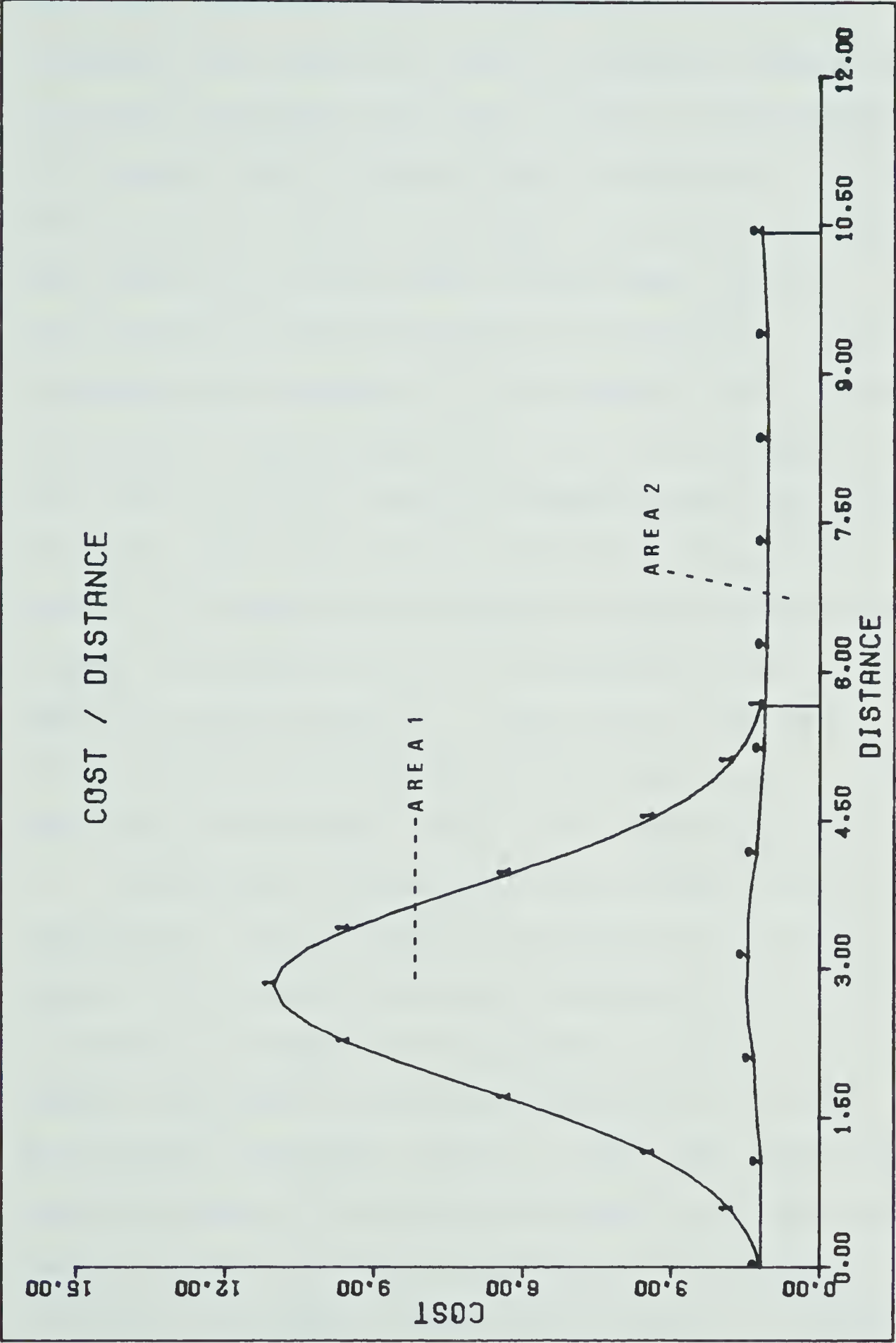


Figure 4 Profile of cost versus distance from one end point
for paths in Figure 3

shown by the symbol "BC" in Figure 3. This base cost represents cost which can be applied as a constant throughout the length of the path, such as those arising from road surfacing, the cost of wire, or even energy usage. The isocost lines do not reflect the base cost. The total cost (area) of a path is derived from two costs; the base cost multiplied by the path distance and the cost contributed by the cost surface. For instance, if costs are in dollars and distance is in kilometers, the direct path 1 is 5.66 km. long and costs \$30.60 at an average cost of \$5.41 per km. Path 2, which avoids the high cost center, is 10.42 km. long with a total cost of \$12.25 giving an average cost of \$1.18 per km. In the cost versus distance plot in Figure 4 the area for path 2 is less than that for path 1 demonstrating that path 2 is less costly than path 1. Even though path 2 is of greater distance than path 1, the lower cost per unit distance results in a lower total cost.

Intuitively, minimum cost paths follow lines of least resistance. The inclination for people to take short cuts supports the least resistance principle. Even if the cost criterion is changed from distance or travel time to one of comfort, the least resistance principle still holds. Mountainous landscapes result in roads that wind through mountain passes. The slope and sinuosity of such roads is a measure of the resistance provided by the landscape. According to Burghardt(1969) initial primitive roads located to inter-connect administrative centers in the most

efficient manner possible. The Banff-Jasper Highway connecting the two towns was a dirt road constrained to locations that allowed for easy river crossings and avoidance of steep gradients. The location, however, was mostly through trees, avoiding the more scenic locations. The upgraded highway, with a higher construction budget, provided a more scenic route. If scenery were considered to be the optimization criterion the new highway could be considered to be more cost effective. Optimum paths within urban environments often relate costs to travel time. Ring roads around urban centers are located to trade off the cost of land acquisition with the ability of the road to ease local traffic congestion. Regardless of the type, many paths result from attempts to optimize some real or perceived cost.

Locating a minimum cost path is not a trivial problem. Intuitive methods can result in non-optimal paths, because the chosen path often accumulates too much cost by being either too long or too short. In addition, optimal paths often trade distance and cost. Sometimes paths traversing a short high cost area are more economical than those taking a longer but less costly route. The reverse may also be true. Only since Werner and Boukidis (1963) have intuitive methods been given theoretical and mathematical expression. The provision of such expressions has not reduced the problem to a trivial one, for their analysis requires the solution of difficult partial differential equations. Modern numerical

methods and computers have allowed workers such as Turner (1971) to analyse applied engineering problems and provide the geographer with the tools to analyse a wide range of optimal path problems.

Chapter 2 Known Solution Methods

2.1 Early mathematical development of path problems

In their fundamental dissertation, What is Mathematics?, Courant and Robbins (1941) demonstrated that early mathematics was developed from the geometry of points, lines, and surfaces. In geometry many maximum and minimum problems concerning lines (paths) have been formulated and the solution of these problems has led to the mathematics used today. The following discussion presents some of the classical path problems known for the last twenty centuries.

The Greeks accepted without proof that the shortest path between two points on a plane is a straight line. Although many theorems in the Euclidean geometry depended on this observation, the Greeks found no reason to doubt this axiom. Perhaps the Pythagorean theorem may have served as an intuitive proof. The Euclidean geometry was not only an academic exercise for the Greeks, but also a tool for explaining natural phenomena. Heron, an Alexandrian living in the first century, showed that the angle of incidence of a reflected light ray was equal to the reflective angle. Courant and Robbins considered this discovery to be "the germ of the theory of geometrical optics".

The contribution towards minimum path theories by problems concerning light rays is extensive. During the late renaissance, Snell (1591-1626) formulated the law of refraction which describes the path location of a light ray

as it passes from one medium to another. His law states that $\sin A / \sin B = u / v$, where A and B are the angles the light ray makes with the perpendicular interface, and u and v are the velocities of the light in the two mediums. Later in the seventeenth century Fermat used the calculus to prove that the refracted path of Snell was indeed the shortest time path through the two media. Fermat extended this analysis to light traveling through many media of infinitely small thicknesses and developed his "principle of geometric optics". Losch (1954) used this refraction law to describe the location of the least cost transportation route in a sea/land interface.

Fermat was not the only mathematician who contributed to the calculus and minimum path problems in the 17th century. The famous "brachistochrone" problem posed by Johann Bernoulli for his "incompetant" older brother John created even more interest in minimum path problems. "Imagine a particle constrained to slide without friction along a certain curve joining a point A to a lower point B. If the particle is allowed to fall under influence of gravity alone, along which such curve will the time required for the descent be least?" (Courant and Robbins, page 379 quoting ACTA ERUDITUM a 17th century scientific journal).

The problems of Fermat and Bernoulli were recognized in their time as not being solvable by differential calculus, which required known functions (or curves) for its operation. Euler and Lagrange (1736-1813) found a general

method for determining paths which minimize some variable(s) such as time in the "brachistochrone" problem or distance in Fermat's principle. Their method, called the calculus of variations, was one of the most important developments for the applied mathematics and can be found in the solution methods for many optical, mechanical, physical, and cartographic problems.

Geodesics have been in use in cartography for many centuries, yet their contribution to spatial analysis has until recently been limited mainly because the calculation of geodesics on surfaces other than a plane or a sphere requires a calculus of variations approach. The geodesic is the shortest path between two points on a surface. Geodesics form the basis for understanding and defining many non-Euclidean geometries. Bernard Riemann (1826-1866) discovered the elliptic geometry, in which the geodesics are straight lines. The earth is an ellipsoid and the great circle curves or geodesics on the surface are defined as straight lines in the elliptical geometry.

2.2 General mathematical formulation

Existing general mathematical formulations for defining minimum cost paths may be separated into two types: first, continuous path formulations requiring continuous cost surfaces and secondly, discrete path formulations which have been developed for both discrete and continuous cost surfaces. Werner and Boukidis (1963) provide the most

general formulation. Their development is incorporated into the following discussion.

2.2.1 Continuous path formulations

For a 2-dimensional geographic plane described by x and y coordinates, the cost at any point $p(x,y)$ is given as:

$C=f(x,y) \dots (1)$. Define a continuous path s between two points on the geographic plane as an infinite number of connected links of length ds , where

$$ds=\sqrt{(dx)^2 + (dy)^2} \dots\dots\dots(2)$$

The total cost T of a path constructed between two points $P_0 (x_0 ,y_0)$ and $P_f (x_f ,y_f)$ is given by:

$$\begin{aligned} T(s) &= \int_{P_0}^{P_f} C(x,y) ds, \\ &= \int_{x_0}^{x_f} C(x,y) \sqrt{1+(dy/dx)^2} \, dx \dots\dots\dots(3) \end{aligned}$$

Werner and Boukidis (1963) recognized that the minimization of (3) requires a calculus of variations approach. Equation (3) can be generalized as:

$$\text{Minimize } T = \int_{x_0}^{x_f} F(x,y,y') dx, \dots\dots\dots(4)$$

where $y' = dy/dx$. The solution to this equation is given by the Euler partial differential equation

$$\partial F/\partial y - d(\partial F/\partial y')/dx = 0. \dots\dots\dots(5)$$

Werner and Boukidis (1963) provide a detailed development.

A more rigorous solution description for the general equation (4) was given by Howard et al. (1968). The coordinate variables are placed in parametric form as $x(t)$ and $y(t)$. The parameter t could be defined as travel time. Then, equation (4) becomes:

Minimize $T = \int_{t_0}^{t_f} F(x(t), y(t), \dot{x}(t), \dot{y}(t)) dt, \dots\dots\dots (6)$

where $\dot{x}(t) = dx/dt$, and $\dot{y}(t) = dy/dt$. The solution of (6) is given by the Euler-Lagrange equations:

$$\frac{d}{dt} \left(\frac{\dot{x} C(x, y)}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right) = \sqrt{\dot{x}^2 + \dot{y}^2} \frac{\partial F}{\partial x}, \dots\dots\dots (7.1)$$

$$\frac{d}{dt} \left(\frac{\dot{y} C(x, y)}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right) = \sqrt{\dot{x}^2 + \dot{y}^2} \frac{\partial F}{\partial y}, \dots\dots\dots (7.2)$$

which are the conditions for an optimal trajectory as developed by Pontryagin (1961). This approach, called the optimum curvature principle by Howard et al. (1968), was developed for the optimum location of highways.

2.2.2 Discrete formulations

Generally not all cost surfaces or paths can be represented in continuous form. Werner and Boukidis (1963) suggest that continuous problems are specific cases of the more general discrete formulation. The integral of equation (3) becomes a summation. The total cost from point $P(x_0, y_0)$

to $P(x_{n+1}, y_{n+1}) = P(x_f, y_f)$ is defined as:

$$T = \sum_{i=0}^{n-1} C(x, y) \sqrt{\Delta x_i^2 + \Delta y_i^2}, \dots\dots\dots (8)$$

where n = number of links, $\Delta x_i = x_{i+1} - x_i$, and $\Delta y_i = y_{i+1} - y_i$. As Δx_i , Δy_i , approaches zero (8) approaches the continuous formulation (3). The path s is defined by the points $P(x_i, y_i)$, $i=1, \dots, n$. This discrete formulation is more general than the continuous one because of the different ways the cost surface can be defined. The cost $C(x, y)$ defined at a point $P(x, y)$ can be represented in a number of ways:

- by a continuous function as defined by (1)
- by contiguous areas of constant cost (cost planes)
- by a matrix of costs representing values at points

$P(x, y)$ in the region of interest. The solution methodology appropriate to minimizing (8) depends on the characteristics of the cost surface.

If the cost surface $C(x, y)$ is smooth the Euler-Lagrange equations (7.1) and (7.2) may be used to find the minimum path. Not all cost surfaces provide for simple analytical solutions to the Euler-Lagrange equations, so numerical techniques as suggested by Howard et al. (1968) must be used. Rankin (1979) developed an "artillary" method to solve the Euler-Lagrange equations. His method utilizes the boundary conditions that require the solution to pass through both end points. By shooting from one point to the other with

different angles and path lengths, and integrating the Euler-Lagrange equations using a 4th order Runge-Kutta algorithm, Rankin's method converges to an extremum path. This path is then examined for a possible minimum.

A general method for finding minimum cost paths on cost surfaces defined by contiguous areas of constant cost has been presented by Werner (1968). He extends Snell's refraction principle and the application of Losch to a path traversing a number of cost planes and proves that a unique set of refraction angles exist for the minimum cost path. Wardrop (1969) also uses the refraction method in calculating optimum paths for continuous cost media.

Minimum path algorithms provide solution paths for cost surfaces represented at discrete points. Each link between two points on the geographic plane is given a value proportional to the cost of connecting the two points. A matrix of costs is constructed, where a node (or element) (i,j) of the matrix provides the cost of the link connecting the point $P(x_i, y_i)$ with $P(x_j, y_j)$. Two types of algorithm are described by Steenbrink; the tree building algorithms, which build the shortest tree from each node to all other nodes, and matrix algorithms, which calculate the shortest paths between all nodes and all other nodes and store the paths in matrix form. Moore (1959) developed the earlier tree building algorithms. Following this, "once through" tree building algorithms were published by Dijkstra (1959) and Whiting and Hillier (1960). Efficient matrix algorithms were

developed by Floyd (1962), Dantzig (1966) and Farbey et al. (1967). One algorithm developed by Murchland (1967) allows for a change in one link without recomputing all the shortest paths. Steenbrink discussed heuristic procedures which provide very good solutions with efficient use of computer resources. One such algorithm developed by Murchland (1967) allows for a change in one link without recomputing all the shortest paths. Steenbrink also mentions the use of dynamic programming as is incorporated into the solution procedures described by Turner (1971). Goodchild(1977) recently developed heuristic procedures to find minimum cost paths through lattices defined in an urban setting.

2.2.3 Alternative Methods

Alternative methods for finding optimum paths have been developed by geographers and engineers in order to solve specific problems within their respective disciplines. Surface transformation methods are the most important of the alternate methods for geographers, because they have a solid basis within cartography. The engineer, who is more interested in practical applications of the optimum path methodology, has developed complete computer systems for the analysis of cost data and feasible paths.

The soap film experiments of the Belgium physicist Plateau (1801-1883) provided general insights in solving calculus of variations problems. Plateau observed that soap

films suspended between surfaces and or edges minimized their areas. Courant and Robbins (1941) demonstrate a wide range of geometrical, topological and variational problems solved by soap film analogies. Steiner's problem and its three-dimensional extensions are easily solved. An experiment to solve the problem of this thesis could also be designed using soap films. A film is suspended between the closed area bounded by the plane, the cost surface, and two perpendicular edges above the end points. After the soap film is stabilized, the contact by the film on the plane should trace the minimum cost path.

Minimum paths for radially symmetric cost surfaces have received recent attention by Angel and Hyman (1972a), and Puu (1978 a,b,c). They have transformed the velocity (cost) surface into another surface where upon the geodesic corresponds to the minimum time (cost) path. Their methods are discussed in Chapter 5.

A systems approach employing a wide range of algorithms has been used for optimal route location by engineers in both North America and Europe. Howard et al. (1968) suggest that their optimum curvature principle can be incorporated within an engineering system such as "TIES" (Total Integrated Engineering System) described by Schureman (1965) or the Massachusetts Institute of Technology ICES (Integrated Civil Engineering System) as described by Ross and Schemakcer (1965). Dynamic programming techniques are used to solve the Euler-Lagrange equations in these systems. More

recent systems such as OECD (Road Research Group in Paris), developed in Paris, and GCARS (Generalized Computer-Aided Route Selection System) developed by Turner and Miles (1971) utilize a combination of minimum cost algorithms and dynamic programming methods. The GCARS system uses the once through algorithm of Whiting and Hillier (1960) in conjunction with an objective function which consists of a "utility surface" calculated with a combination of trend surfaces representing local cost factors and the distance calculated from a network of interconnected links.

2.3 Discussion

Until now, the existing methods have been mentioned, but no attempt has been made to evaluate them as to their suitability to solve theoretical and empirical geographic problems. The following discussion considers the methodology chosen for this thesis. A common set of criteria which can be used to evaluate these methods is difficult to obtain, because the existing methods are tailored to solve specific types of problem.

The first of the methodologies is contained within the mathematical development given by Werner and Boukidis (1963) and Howard et al. (1968). Although the Euler-Lagrange equations (7.1) and (7.2) provide a good theoretical definition of the location of the minimum cost path, the exact solution of these equations is generally very difficult. This is supported by Courant and Robbins (1941).

"It is usually very difficult, and sometimes impossible, to solve variational problems explicitly in terms of formulas or geometrical constructions involving simple elements. Instead, one is often satisfied with merely proving the existence of a solution under certain conditions and afterwards investigating properties of the solution. In many cases, when such an existence proof turns out to be more or less difficult, it is stimulating to realize the mathematical conditions of the problem by corresponding physical devices, or rather, to consider the mathematical problem as an interpretation of a physical phenomenon. The existence of the physical phenomenon then represents the solution of the

mathematical problem." (Courant and Robbins, page 386)

Howard et al. (1968) provide a family of solutions to the Euler-Lagrange equations for an exponential cost model by finding initial-value conditions that satisfy the equations. They suggest that solution conditions for the two-point boundary value problem are very much harder to find and that numerical integration of the objective function (6) is an easier problem. Generally, however, the solution of these equations for a geographic application is too tedious. The use of soap film experiments to provide solutions to these equations also has practical problems, because a physical representation of the cost surface is difficult to construct. If constraints are placed on path location, new Euler-Lagrange equations may have to be developed from first

principles. Rankin (1979) mentions that considerable modification to his method must be made before he can incorporate constraints into his numerical method. The lack of literature exploring this problem area indicates the difficulty with this purely theoretical approach.

The representation of both continuous and discrete cost surfaces by many planes of constant cost values, as suggested by Werner (1968) demands the acceptance of two assumptions. Firstly, Werner (1968) assumes that the geographical cost surface may be adequately represented by a set of these contiguous cost planes. Secondly, he assumes that the refracted path at the interface between two cost planes is acceptable and realistic for many types of paths. For paths representing roads through landscapes with small variations in cost the method is proven to be adequate by Werner and Boukidis(1963). However, some cost surfaces representing travel time within an urban environment, may be very difficult to approximate with constant cost planes. The cost surface may have very steep gradients and if the number of cost planes representing this surface is small, the refraction angle of the path may be very large. By increasing the number of cost planes the path may locate in a different region. Werner's method also requires that the set of planes to be traversed by the minimum cost path be known before the refraction angles can be calculated. This requirement may be met by the solution methods proposed in this thesis. In addition Werner suggests that constraints

must be considered, but, he does not incorporate them into his solution methodology. Werner concludes that his particular "model provides only basic concepts on which to develop a theoretical approach".

The most general of all solution methodologies are found in minimum path algorithms. Once the surface is represented by point to point costs a variety of algorithms can be used to calculate the minimum cost paths. The use of these algorithms by engineers indicates their advantage in being suitable for solving practical problems. The disadvantages of using minimum path algorithms are threefold. Firstly, for very large problems the computation costs of some algorithms are unreasonable, even though a variety of heuristic techniques can be used to provide efficiency (see Steenbrink (1974), Chapter 7). Secondly, the solution vector is restricted to the grid points defining the cost surface, hence, the minimum cost path is no more precise than the density of the grid. The third disadvantage concerns constraining the path to certain criteria such as slope, radius of curvature, or other factors. Either the minimum path algorithm or the cost matrix must be modified to incorporate this type of constraint.

The systems TIES and GCARS have been used to resolve many empirical engineering problems concerning minimum cost paths. The GCARS method uses the minimum path algorithms discussed above and mentions the same concern for computational efficiency with large grids. The TIES system

uses minimum path algorithms in conjunction with numerical integration of the cost integral to define alternative minimum cost paths. The expansion of these systems into a general framework which is suitable for theoretical geographic application is a software evaluation problem beyond the scope of this thesis.

2.4 The mathematical optimization formulation

Aside from the GCARS system described by Turner (1971), there have been very few attempts made to use a mathematical optimization approach to solve the basic minimum cost path problem. Steenbrink (1974) suggested that a discrete formulation such as presented by an equation similar to (3) can be very easily solved using dynamic programming if the path can be represented as a sequence of connected points. The use of mathematical optimization to solve minimum cost path problems especially related to road location was discussed in the two symposia of the Planning and Transport Research and Computation (1969 and 1971) and initially by Werner and Boukidis (1963).

A mathematical optimization formulation of the minimum path is given by:

Minimize $F = \sum_{i=0}^{n-1} C(P_i) D_i, \dots\dots\dots (9)$

where $C(P_i) = C(x_i, y_i)$ and $D_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$, and n equals the number of links connecting $P(x_0, y_0)$ and $P(x_f, y_f)$.

Equation (9) is known as an objective function, the meaning of which can be expressed as: "which points $P_i, i=1, \dots, n$ minimize the value of F (or the cost)". A general treatment of this objective function is given in the next chapter.

The use of mathematical optimization has merit. By minimizing the discrete version of equation (3), there is no need to solve the Euler-Lagrange equations (7.1) and (7.2). The objective function is flexible. The general cost surface $C(x,y)$ may contain any combination of cost surfaces as long as each surface is continuously differentiable for every point (x,y) . The path defined by n links can also be modified to include costs dependent on only the path. Both linear and nonlinear constraints imposed on the path can be dealt with according to well developed optimization methodology. Finally, different objective functions can be developed to accommodate a wide variety of path minimization problems.

3.1 Mathematical optimization

The minimization of the objective function (9) may be achieved through the use of non-linear mathematical optimization techniques. A general introduction to mathematical optimization can be found in Aoki(1971), Steenbrink(1974), or in a variety of text books. Steenbrink formulates the general mathematical optimization problem as follows:

"It is desired to determine values for n variables x_1, \dots, x_n in such a way that the value of a function of these variables ($F(x_1, \dots, x_n)$) is as large or as small as possible. The variables x_1, \dots, x_n are called decision variables or instrument variables. The function $F(x_1, \dots, x_n)$ is called the objective function. The objective function must be maximized or minimized.

Moreover, there (may) exist certain relationships between the decision variables and/or the decision variables or functions of the decision variables which must satisfy some inequalities or equalities. These are called constraints:..."

The minimization problem is generally formulated as:
Minimize $F(x_1, \dots, x_n)$(10)
subject to $g_j(x_1, \dots, x_n) \leq 0, j=1,nc,$

and $h_k(x_1, \dots, x_n) = 0, k=1, neq; \dots\dots\dots(11)$

where nc equals the number of inequality constraints and neq equals the number of equality constraints.

The problem is expressed in vector notation as:

minimize $F(X)$, s.t. $G(X) \leq 0, H(X) = 0, \dots\dots\dots(12)$

where $X = x_1, \dots, x_n$. The conditions for minimizing this objective function are seperated into unconstrained and constrained types.

The conditions for deriving a minimum of an unconstrained objective function are now examined. Define X^* as an optimal vector of decision variables that minimizes (10). The necessary conditions for the minimum X^* require that every partial derivative of F with respect to $x_i, i=1, n$ be equal to zero. The vector of partial derivatives is called the gradient and is given by:

$\nabla F(X) = (\partial F / \partial x_1, \dots, \partial F / \partial x_n). \dots\dots\dots(13)$

For a convex F , X^* is a local and a global minimum if and only if the gradient is equal to zero. F is convex if for any X_1, X_2

$F(\lambda X_1 + (1-\lambda)X_2) \leq \lambda F(X_1) + (1-\lambda)F(X_2),$ if $0 \leq \lambda \leq 1$.

If F is strictly convex and the gradient is equal to zero, then X^* is a unique global minimum. F is strictly convex if for any X_1, X_2

$F(\lambda X_1 + (1-\lambda)X_2) < \lambda F(X_1) + (1-\lambda)F(X_2),$ if $0 < \lambda < 1$.

The sufficient conditions for a local minimum can be tested by deriving the Hessian matrix of dimension $(n \times n)$. This matrix is calculated by taking all the partial

derivatives of the gradient at X^* which are given by:

$$\text{HESS}(X^*) = \partial^2 F / \partial x_i \partial x_j, \dots\dots\dots (14)$$

where $i=1,n$ and $j=1,n$. A local minimum can be declared if the Hessian matrix is positive definite. A symmetric matrix is positive definite if and only if all eigenvalues are greater than zero. A positive semidefinite matrix requires all eigenvalues to be greater or equal to zero. If the Hessian matrix is positive semidefinite at X^* , no decision on whether X^* is a local minimum can be made. Instead the eigenvalues of the Hessians evaluated in the neighborhood of X^* must be examined. If the eigenvalues of these Hessians are all non-negative, then a local minimum can still be claimed. A more complete discussion of the conditions required for local and global minimization is found in Steenbrink(1974), Aoki(1971), and Hauer(1974).

For general minimization problems with constraints, the objective function (10) and constraints (11) are combined with the use of Lagrangian multipliers to give the unconstrained formulation:

$$\text{minimize } K(X,L,U) = F(X) + L^T G(X) + U^T H(X), \dots\dots\dots (15)$$

where $L=l_1, \dots, l_{nc}$ and $U=u_1, \dots, u_{neq}$. Given that X^* is a local minimum which satisfies the equality and inequality constraints given by (11) and at X^* a suitable constraint qualification holds, then there exist $l_i \geq 0, i=1,nc$, such that the following conditions hold:

$$K_x(X^*,L,U) = F_x(X^*) + L^T G_x(X^*) + U^T H_x(X^*) = 0,$$

and $L^T G(X^*)=0$,(16)

where L and U are as defined in (15). These equations are known as the Kuhn-Tucker conditions defined by Kuhn and Tucker (1951).

Aoki(1971) classified the methods of minimizing or maximizing an objective function into three categories:

- (1) Methods using only the functional values, called direct methods.
- (2) Methods making use of the first-order derivatives as well.
- (3) Methods which also require knowledge of second-order derivatives. Aoki(1971).

Generally, the solution methods used depend on the characteristics of the objective function. Direct methods are used for objective functions for which the first and second derivatives are difficult to calculate. First-order derivative methods, commonly known as gradient methods, offer the most flexibility for solving problems which have difficult solutions and imposed constraints. Methods utilizing second derivatives require an objective function and constraints with existing second-derivatives. Most methods locate stationary points and the convexity conditions must be examined at these points before local or global minima can be claimed.

Several optimization techniques using gradient methods have been programmed by Hauer(1974). These methods utilize the first order derivatives of the objective function in an iterative procedure. An initial approximate solution X and a search direction equal to the negative reciprocal of the gradient is given to the program. The next approximate solution X is found with the recursion $X(k+1)=X(k)+t(k)D(k)$, $k=0,1,2,3,\dots$; where k is the iteration, D is the feasible search direction and t is a scalar determining the step size. After each iteration a new feasible search direction is determined from the gradient at $X(k)$. The procedure stops at a stationary point if the following conditions occur:

- (1) if the gradient becomes sufficiently close to zero for $X(k)$ in an unconstrained region
- (2) if any search direction results in a non-decrease in the value of the objective function
- (3) if no significant decrease in F can be achieved in a fixed number of iterations.

For constrained minimization the Kuhn-Tucker conditions at the stationary point are examined. If the conditions (16) are satisfied, and if the objective function and the inequality constraints are convex, while the equality constraints are affine, the stationary point is declared a global minimum at the constraints. For unconstrained

minimization, the gradient and the Hessian at the stationary point are examined for a local minimum.

The advantages of using the gradient methods for our application can be summarized as follows:

- (1) The objective function is smooth and convergence occurs in a few iterations.
- (2) The computer algorithms available concentrate on gradient methods.
- (3) Constraints can be incorporated into the solution procedures.

A more complete description of these methods is found in Aoki(1971) and Hauer(1974).

3.2 Modelling the minimum cost path as an objective function

The objective function (9) $F = \sum_{i=0}^{n-1} C(P_i) D_i$ may clearly represent the theoretical problem, but it is not a particularly useful function for optimization. The locational variables P_1, \dots, P_n as defined in section 2.4 are too general and must be represented by a set of decision variables which can be manipulated. There are many possible ways of combining the x and y coordinates of the locational variables into this set and not all methods provide good solutions of the objective function. This section addresses the problem of choosing a set of decision variables which provide optimum paths and at the same time satisfy the

operational requirements of the solution methodology.

3.2.1 The two link path and the objective function

The technique of isodapanes developed by Weber (1909) may be used to demonstrate the objective function. Isodapanes are lines of equal total cost from two fixed points on the plane to a variable third point. If the objective function (9) is restricted to a single locational variable, then the contours of the function correspond to Weber's isodapanes. Isodapanes were used by Smith (1966) to find locations which minimize the total transportation cost of raw materials and finished products. Similarly, the contours can be used as a graphical tool to find the location on the plane, where the objective function is at a minimum. The general form of the objective function with a single locational variable $P(x,y)$ is given by:

$$F(x,y)=Co(x,y)\sqrt{(x-x_0)^2+(y-y_0)^2} + Cf(x,y)\sqrt{(x-x_f)^2+(y-y_f)^2}, \dots\dots\dots(17)$$

where $Co(x,y)$ is the average cost above the first link and $Cf(x,y)$ is the average cost above the last link in the path. This average cost is explained in the Section 3.2.2. The decision variables of this objective function are the x and y coordinates of the third point. Objective functions for a homogeneous and an inhomogeneous cost surface are now considered.

Figure 5 presents the elliptical contours (isodapanes)

for an objective function with a homogeneous cost surface $C(x,y)=5.0$. This cost surface is plotted in Figure 1. A point anywhere on the straight line between the two end points located at $P(-2,2)$ and $P(2,-2)$ provides an optimum solution for the two link path. Substituting the values in (17) the objective function becomes:

$$F(x,y)=5.0(\sqrt{(x+2)^2+(y-2)^2} + \sqrt{(x-2)^2+(y+2)^2}). \dots\dots\dots(18)$$

Using the property of distance and the triangular inequality, it can be shown that this objective function is convex. The objective function is the sum of two distance functions multiplied by a constant. The distance function is convex and the sum of two distance functions multiplied by a constant is also convex indicating a convex objective function. The particular minimum found for the objective function in Figure 5 is at $P(x,y)=(0.3696,-0.3696)$.

Figure 6 demonstrates the contours of the two link objective function for the cost surface shown in Figure 3. The cost surface is given by:

$$C(x,y)=10.0\exp(-0.2x^2-0.8y^2)+1.0. \dots\dots\dots(19)$$

The base cost is incorporated into this cost surface and is given the value of 1.0. The objective function now becomes

$$F(x,y)=Co(x,y)(\sqrt{(x+2)^2+(y-2)^2}) + Cf(x,y)(\sqrt{(x-2)^2+(y+2)^2}), \dots\dots\dots(20)$$

where $Co(x,y)$ and $Cf(x,y)$ are the average costs of the first and last link evaluated using the cost surface described by

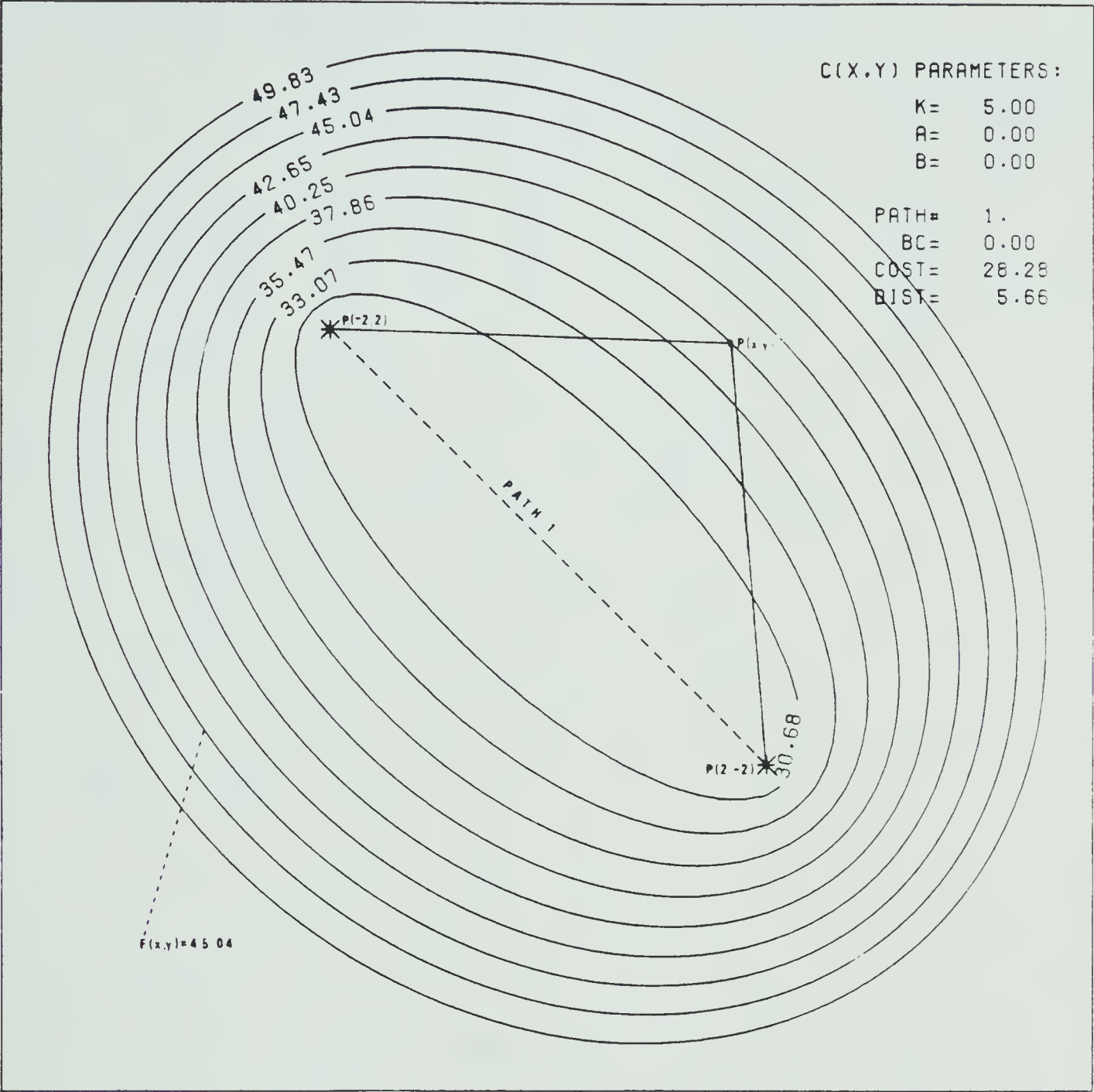


Figure 5 Contours and two link path for objective function (18)

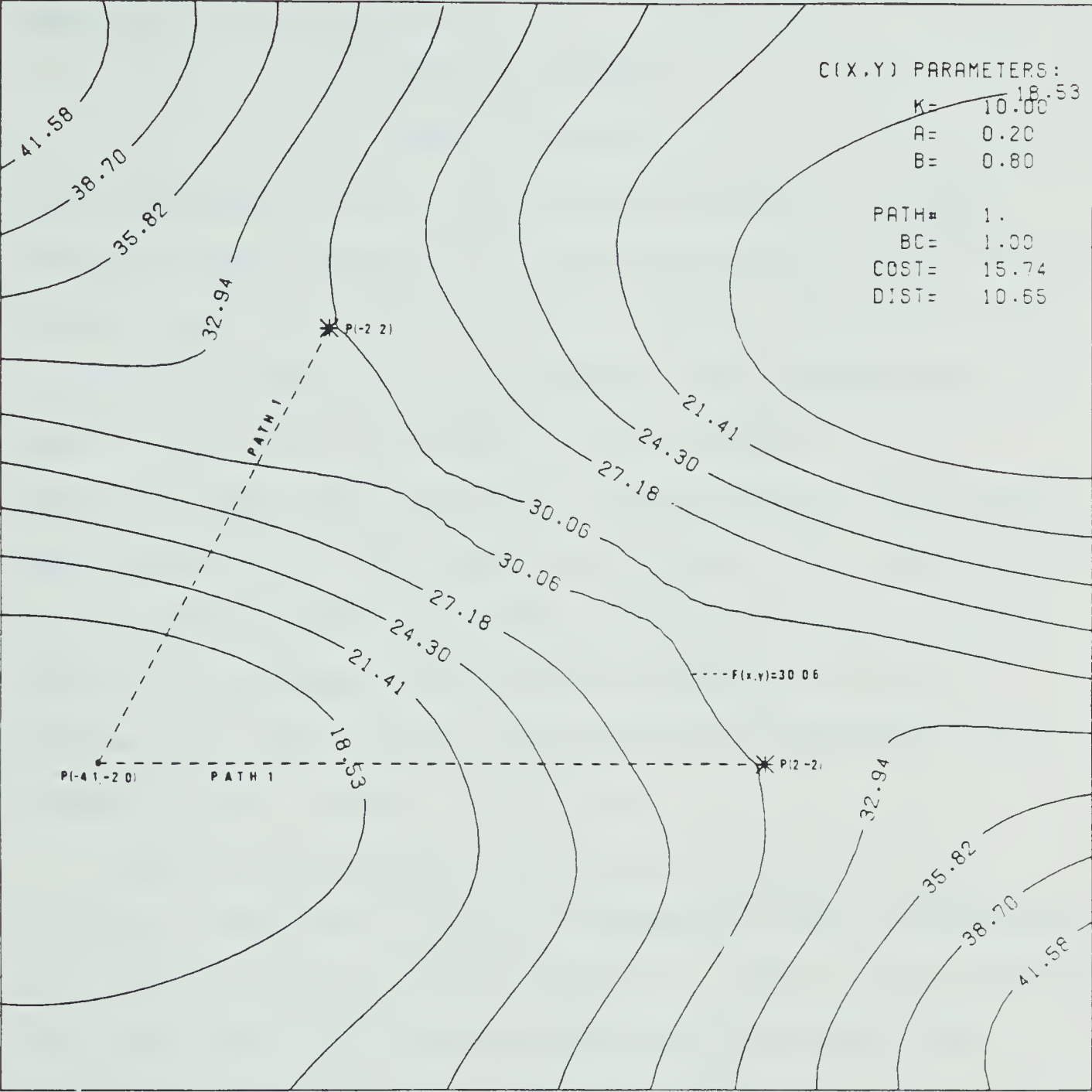


Figure 6 Contours and two link path for objective function (20)

(19) above. The fixed end points are located at $P(-2,2)$ and $P(2,-2)$. A minimum value of the objective function is found at $P(x,y)=(-4.1129,-2.0154)$. The Hessian evaluated numerically at this point is:

$$\begin{matrix} 0.8174378 & 0.1114070 \\ 0.1114070 & 2.2476149. \end{matrix}$$

The eigenvalues for this Hessian are 0.8088115, 2.256241, indicating that the matrix is positive definite and that the minimum found is a local minimum. Because the cost surface is symmetric about the axes, another local minimum can be found in the other pit shown by the contours. It is speculated that these two local minima belong to the set of global minima for this objective function. For large $|x|$ or $|y|$ the cost surface approaches the value of 1.0. In this region of the plane this particular objective function behaves like the distance function with increasing value. A minimum in this region is unlikely.

Clearly, the optimal solutions for the two link objective functions with both homogeneous and inhomogeneous cost surfaces can be found graphically. Such is not the case for paths with more than one locational variable. The objective function surface for these paths extends into a space of more than the three dimensions which is not easily visualized. The contouring of such surfaces is virtually impossible and the numerical solution techniques outlined above are required.

3.2.2 Variable x and y decision variables

An obvious choice for the decision variables are the x and y coordinates of the locational variables as described in the two link method above. By allowing each (x,y) pair to locate freely on the geographic plane one can argue that some location of such pairs should satisfy the objective function and produce a minimum cost path. The objective function using these decision variables is given by:

$$F(X,Y) = \sum_{i=0}^{n-1} (Ca(x_i, y_i) + BC) \left(\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \right), \dots (21)$$

where n is the number of links in the path, x and y are the coordinates, $X=x_1, \dots, x_{n-1}$, $Y=y_1, \dots, y_{n-1}$, and BC is the base cost. Ca equals the average cost of link i and is given by:

$\sum_{j=1}^m C(x_{ij}, y_{ij})/m$, where $x_{ij} = (x_{i+1} - x_i)j/m$ and $y_{ij} = (y_{i+1} - y_i)j/m$. For a n link path 2n decision variables are required.

The x,y decision variable method is illustrated in Figure 7. The cost surface and end points are the same as for Figure 6. The number of links n in this path is 8 and the cost of each link is averaged at m=5. The number of decision variables is 14. The solution found in Figure 7 is a local minimum because the eigenvalues of the Hessian are all positive.

3.2.3 The variable r and theta method

A second method for defining the decision variables of

C(X,Y) PARAMETERS:

K= 10.00
A= 0.20
B= 0.80

PATH# 1.
BC= 1.00
COST= 12.35
DIST= 10.44

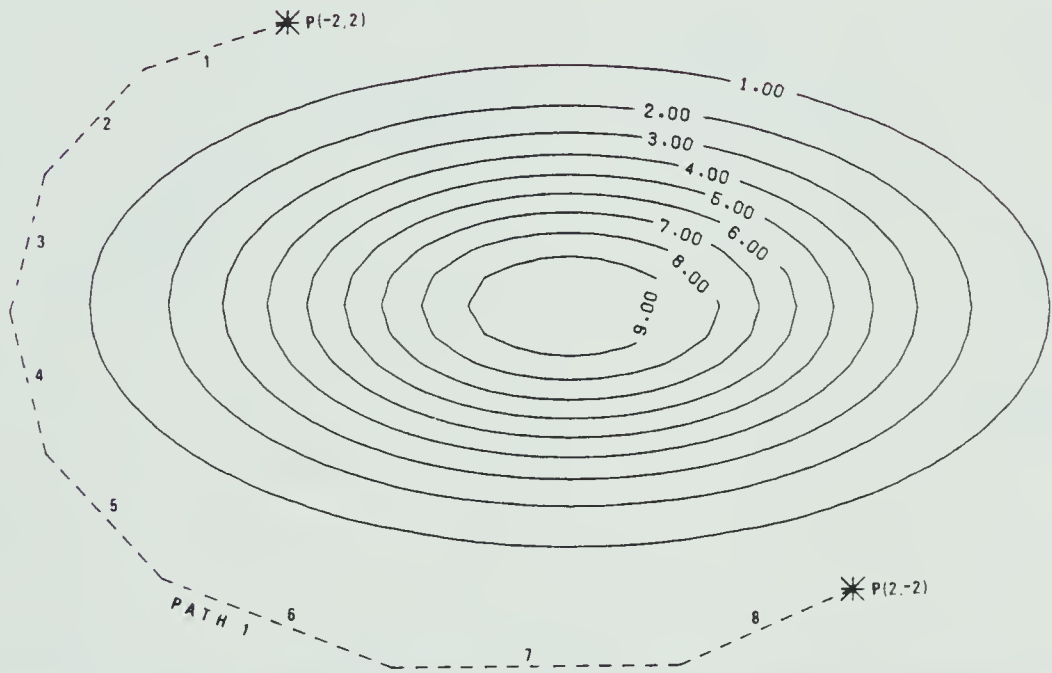


Figure 7 A path calculated using the x,y method

the objective function is proposed. Let the path be defined by a fixed number of equal but variable length links. Let each of these links vary their orientation on the x,y plane. Then the decision variables are the angles "theta" each link makes with the x-axis, and the length "r" of each and every link. Using this method the path is represented by the following schema:

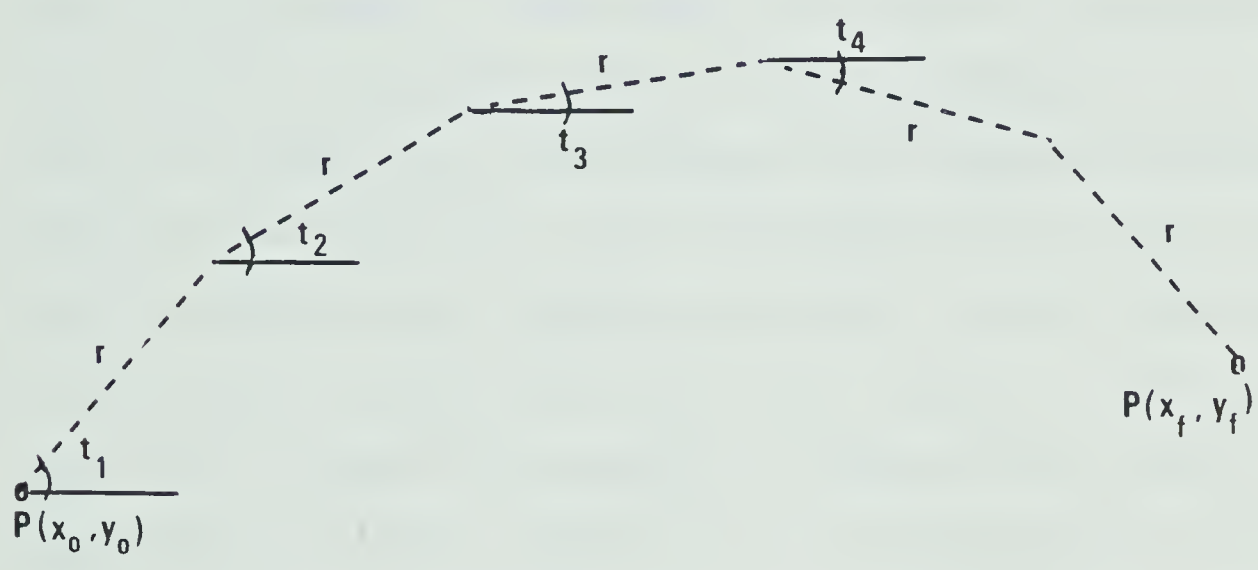


Figure 8 A five link path with equal but variable link length

The decision variables for this representation are the angles theta t_1, \dots, t_4 and the link length r . The coordinates x and y of the locational variables $P(x,y)$ are given by:

$$\begin{aligned} x_i &= x_0 + r \sum_{j=1}^i \cos t_j, \\ y_i &= y_0 + r \sum_{j=1}^i \sin t_j, \dots\dots\dots (22) \end{aligned}$$

where n equals the number of links.

In order to incorporate the decision variables t , and r into the objective function special attention is given to the relationship between the last link and r . N steps of r and respective angles t defines the location of all links of the path, including the last. If the path is to reach the end point with last link of length r exactly, then r must be set to the length of the last link. This is achieved through the use of a quadratic penalty function added to the objective function. The minimum of the squared difference between the length of r and the length of the last link occurs when the two lengths are equal, and the contribution of the penalty function to the total value of the objective function is zero. This method is theoretically sound, but it has a few practical problems. As the optimization proceeds, the length of r is acted upon by two forces. The optimum region of the path requires one length of r and the penalty function requires another. When the path is located near the optimum, small changes in r required to zero the penalty function do not cause large changes in the value of the objective function and often a stationary point resulted. Fortunately, the solution at this stationary point is close to a local minimum because the gradient is near zero.

The variable link length method discussed above is incorporated into the following objective function:

$$F(T,r)=\sum_{i=0}^{n-1}(Ca(x_i,y_i)+BC)r + w(\sqrt{(x_f-x_{n-1})^2 + (y_f-y_{n-1})^2} - r)^2 , \dots\dots\dots(23)$$

where n is the number of links in the path, Ca is as defined in Section 3.2.2., r is the link length, $T=t_1, \dots, t_{n-1}$ are the angles each point makes with the x-axis (see Figure 8), BC is base cost of each link and w is a constant weight. A further explanation of the objective function (23) is required. The last term is actually a penalty function added to the cost of the path. For a large enough value of w, the length of r will be set equal to the length of the last link. In order that the penalty function does not contribute to the cost part of the objective function, the last link and r must have the same value. Through a series of experiments a weight of w=10.0 was chosen. This weight brought r close enough to the length of the last link such that at the minimum, the penalty function contributed very little to the value of the objective function. This weight of w is used for all future calculations for the variable r objective function.

Figure 9 illustrates a path calculated using the variable r and theta method. The problem posed is the same as that for Figure 7. For the 8 link path, the variable r and theta method requires 8 decision variables. The path demonstrates the equal length of links. The path in Figure 9 is declared a local minimum.



Figure 9 A path calculated using the r and theta method

3.2.4 A comparison of the two methods

The variable x,y , and r and θ methods can best be compared by two numerical experiments. The first experiment is designed to compare the location of the path calculated using each method. Figure 10 illustrates that the paths calculated with each method are located in the same region of the plane. The only noticeable difference in this figure is the cost contributed to each path. The variable x,y method provides a path with 0.02 less cost. This difference in cost can be explained by the path shown in Figure 7, which is calculated using the x,y method. In the highly curved regions of the path the link length is shorter. In the more straight regions, the link length increases. As a result the x,y method produces paths which are smoother in their curved extents. This smoothness may result in more cost effective paths. The second experiment illustrates the performance of each method for calculating paths with increasing number of links. Figure 11 clearly indicates that the cost of the path for each method converges as the number of links increase. At 10 links, for example, the cost becomes virtually identical. However, the number of iterations required to calculate the path for increasing number of links is radically different. Figure 12 shows that for the variable x,y method, the number of iterations increases at 10 to the power of a constant times the number of links. The variable r and θ method converges at a

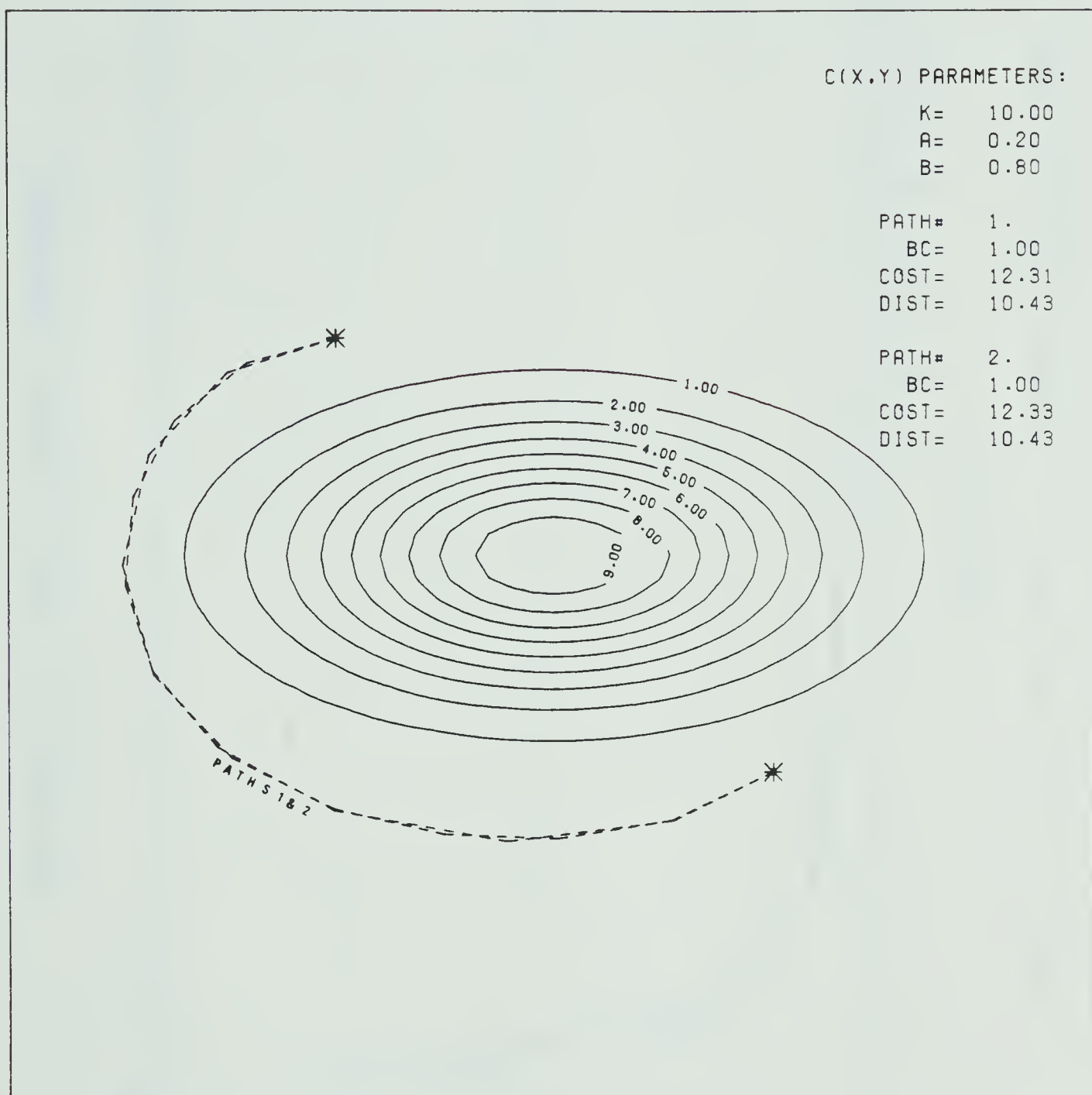


Figure 10 Paths calculated using the x,y and r, theta methods

METHODS X, Y AND R, THETA COMPARED

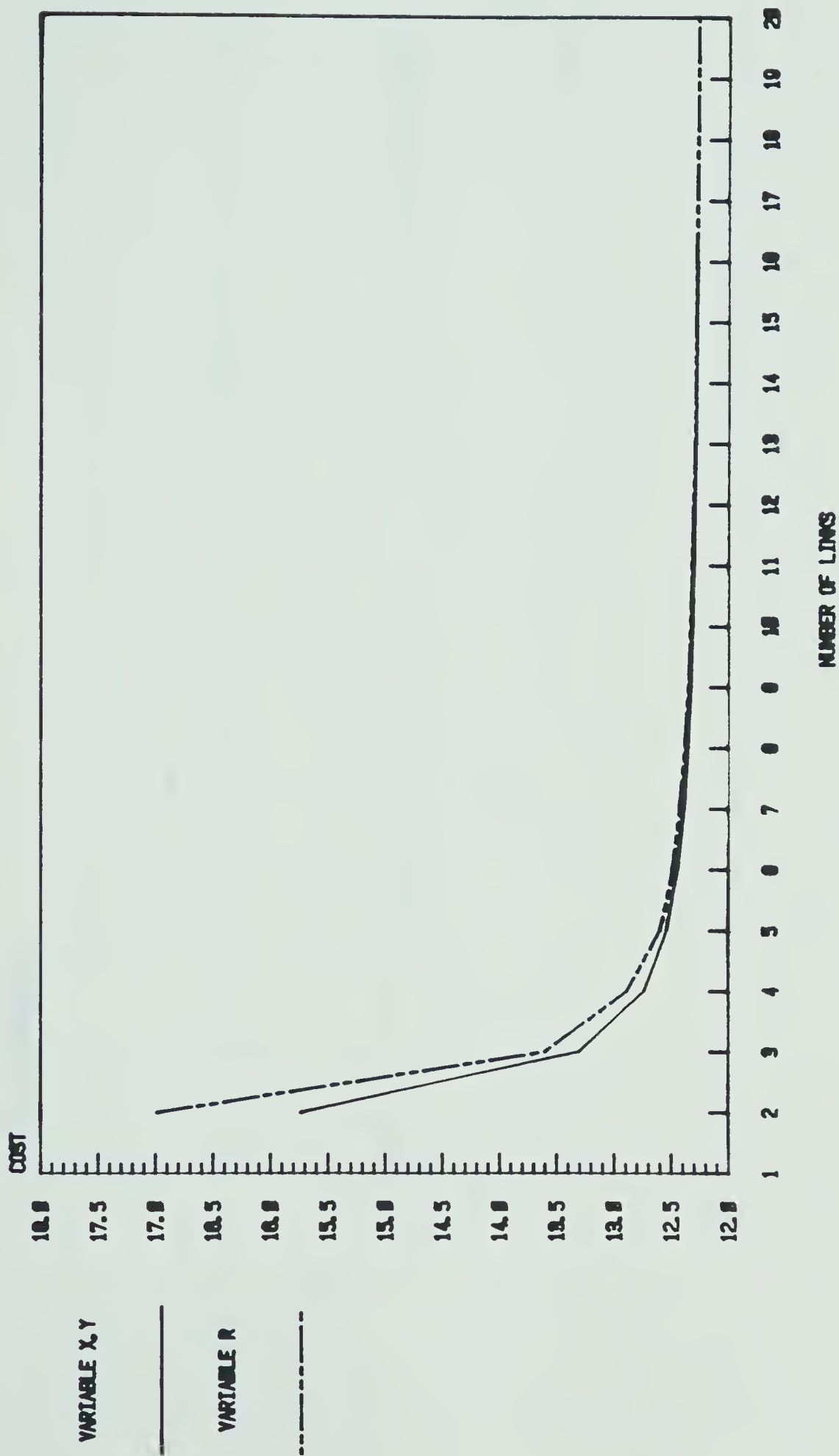


Figure 11 Cost calculated at different number of links

METHODS X, Y AND R, THETA COMPARED

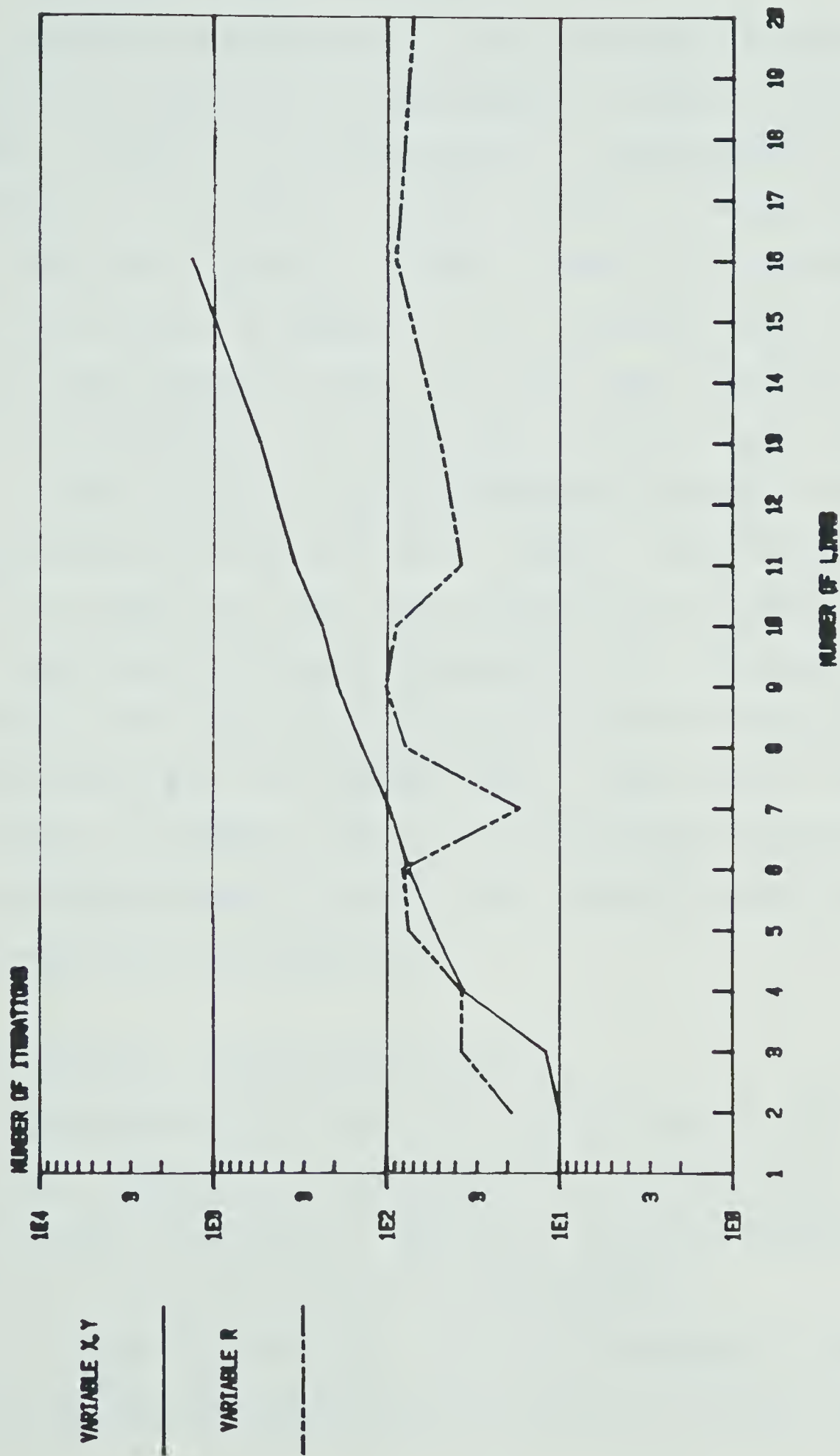


Figure 12 Number of iterations required for convergence at different number of links

relatively constant number of iterations regardless of how many links were chosen in the path.

A possible explanation for the difference in the number of iterations required for convergence in each of the two methods is found in the relationship of the decision variables to one another. In the variable x, y method, the x and y for each locational variable effects only the adjacent links. As a result a change in x or y at each iteration effects the location of only two links. Many iterations are required to transmit the change down all the links of the path. In the variable r and θ method a change in the orientation of the first link for example, effects the location of the last link. This observation is supported by the relationship defined in formulation (22). Because the variable r and θ method provides virtually the same solution for paths with greater than 9 links and for fewer iterations, it appears that it is the preferred method for calculating minimum cost paths. Other numerical experiments have supported this observation.

3.2.5 Numerical integration analysis

The objective functions (21) and (23) are in part numerical integrations. Confidence in it's ability to approximate a continuous integration is demonstrated by Figure 13. The cost surface is represented by $C(x, y) = 10.0 \exp(-0.2x^2 - 0.8y^2)$. Four paths between the same end points were calculated with 2, 4, 8 and 16 links. The

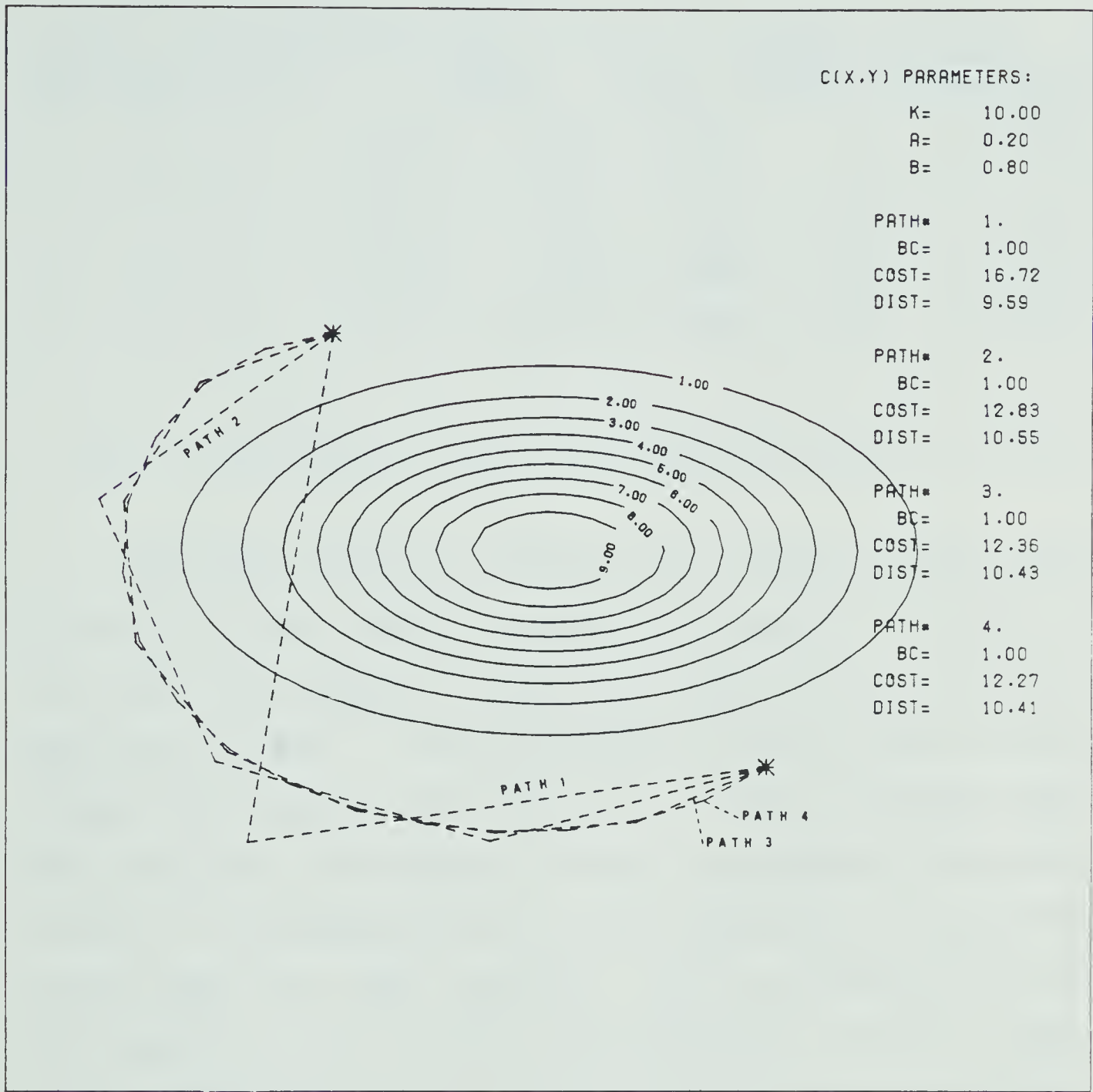


Figure 13 Paths calculated with 2, 4, 8, and 16 links

distance, cost, F value, number of iterations and CPU time for each path were calculated and presented in Table 1.

Number of links	Path #	Distance	Cost (area)	F value	Number of iterations	CPU seconds
2	1	9.594	16.724	16.841	7	0.054
4	2	10.553	12.826	12.832	19	0.140
8	3	10.434	12.359	12.360	24	0.309
16	4	10.411	12.273	12.273	25	0.491

Table 1 Comparison of paths with varying number of links

The paths plotted in Figure 13 all locate within close proximity of each other. The paths with greater than two links are very similar in both cost and distance as can be observed in Table 1. The difference in the precision of the integration between paths with 8 or 16 links indicates that the choice for the number of links is a subjective decision. Figure 11 illustrates that the cost approaches an asymptote as the number of links increase. A path calculated with 10 links appears to fall within the asymptotic portion of the curve. The distance and cost of the path calculating using Rankin’s (1979) method was 10.404 and 12.247 respectively. This compares favourably with the 16 link path.

3.2.6 Other solution criteria

A solution to the objective functions (21) and (23) must satisfy other numeric and graphic criteria before it can be accepted as a reasonable solution. The numeric criteria requires the same value of F for paths with the same optimum solution and different starting solutions; and the same value of F for symmetrical solutions. Graphic criteria is found in the location of the path. The numeric criteria may indicate an optimal solution, but the solution may not make any graphic sense. All solutions must be visually examined. Figure 14 shows two different starting solutions which converge to the same optimum solution. The cost surface is given by $C(x,y)=10.0\exp(-0.2x^2-0.8y^2)$. The paths plotted are path 1 with optimum solution path 2 and path 3 with optimum path 4. Path 2 and path 4 are identical as indicated by the values of F and path distance.

Sometimes there may be more than one equivalent solution to a particular problem. Intuitively, if both the cost function and the beginning and end points are symmetric about an axis, there should be two symmetrical solutions. Figure 15 demonstrates two virtually identical solutions for a cost surface given by $C(x,y)=10.0\exp(-0.9x^2-0.1y^2)$.

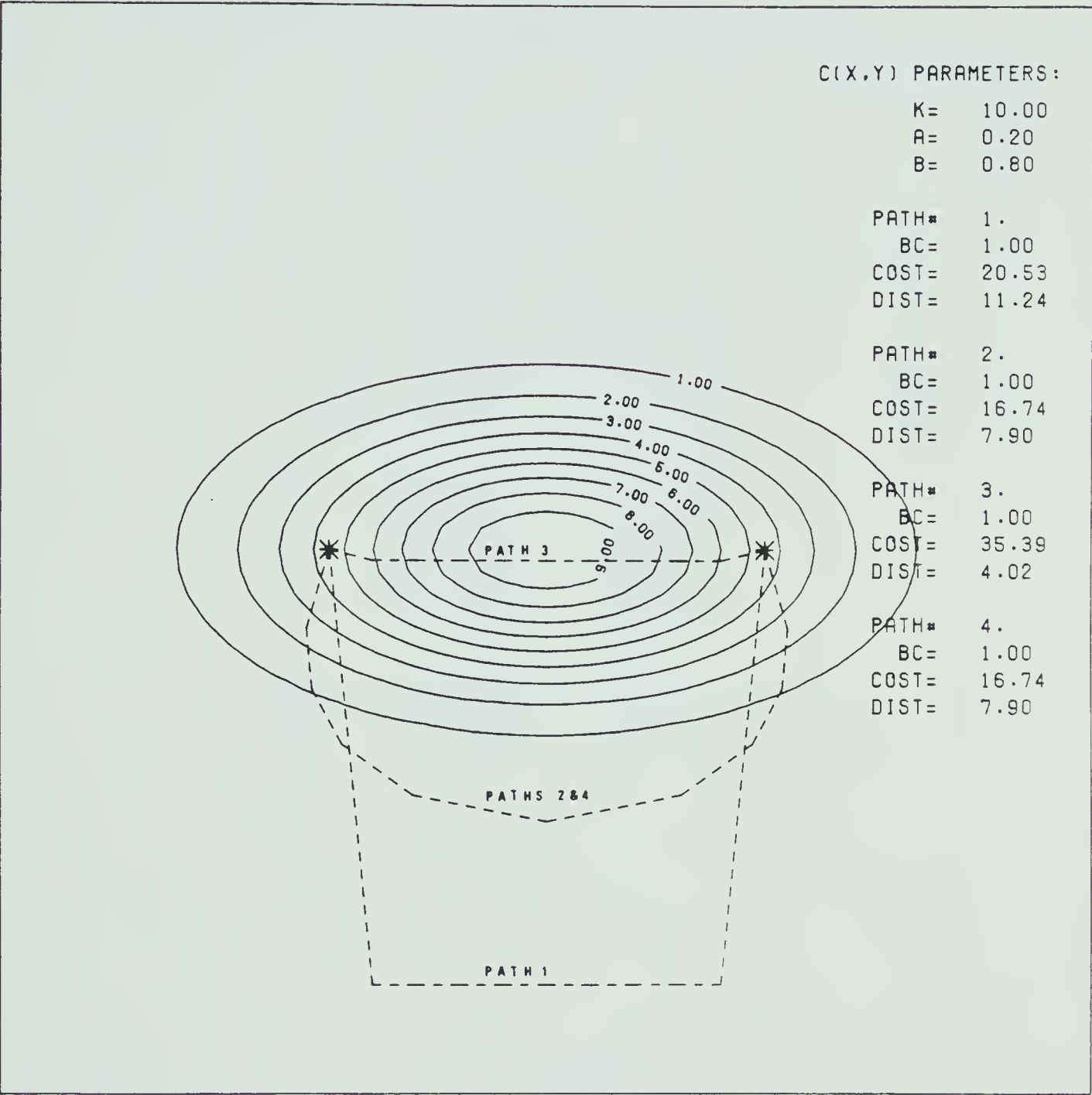


Figure 14 Two initial solutions - one optimum solution

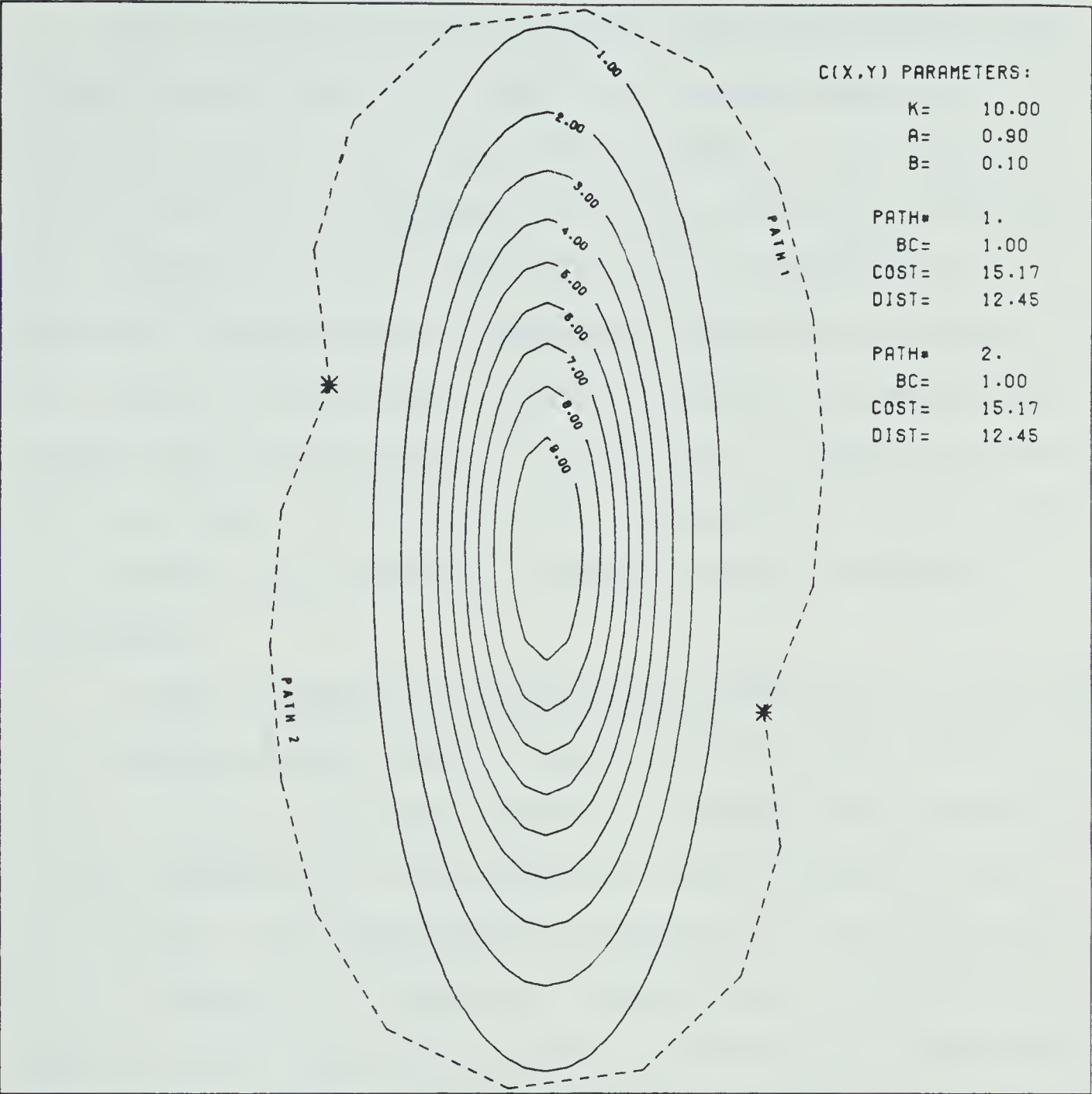


Figure 15 Two symmetrical solutions

3.3 Constrained solution paths

Not all minimum path problems can be expressed by the objective functions (21) and (23). Some minimum path problems require that the solution path be restricted to certain areas of the geographic plane either by natural or man imposed barriers. Other paths may be subjected to penalties imposed by the geographic space they traverse. In some cases such problems can be formulated by subjecting these objective functions to constraints. These constraints are expressed in two ways: first, through the use of linear constraints, and secondly, through the use of penalty functions.

Linear constraints formulated by equation (15) are used for finding minimum paths constrained by linear barriers on the x,y plane. For each locational variable there exists a region beyond which the constraints are violated. Inside this region the minimization proceeds as if no constraints exist. Figure 25 in Chapter 4 demonstrates a constrained path. For efficient optimization, starting solutions within the feasible region of the surface are preferred. These are known as feasible solutions. Hauer (1974) expresses linear inequality constraints as a linear system described by $G(X)=AX + B \leq 0$. The matrix A and vector B , and the decision variables X representing the x,y coordinates of the path, can force the path to be restricted to certain regions of

the plane.

Constraining the path by adding an artificially high cost to the cost surface in the region of interest may be achieved through the use of penalty functions. This technique suggested by Werner (1968) is concerned with using a function to force a particular locational variable to locate outside the range of influence. The use of penalty functions can be incorporated into any cost surface and no special optimization procedure is required.

Chapter 4 Examples of Minimum Cost Paths

The location of minimum cost paths is as conjectured in previous chapters generally not achieved through intuitive insights. By presenting examples of minimum cost paths this chapter demonstrates two points. First, that the problem of where to locate minimum cost paths is not trivial and second, that the method can be used to solve different types of minimum cost path problems.

The examples presented in this chapter are divided into three sections. Section 4.1 presents a variety of problems exploring some general relationships between the cost surface, the end points and the minimum cost path. The second section concentrates on finding minimum cost paths for cost surfaces which are derived from the current literature. By providing identical examples, the methods used in this thesis are supported. Section 4.3 explores the effect of simple constraints on path location. The examples which use constraints further support the use of mathematical optimization. In all examples presented the number of links was kept at 10 and the penalty weight w of the objective function was set at 10.0.

4.1 Variations on one cost surface

During the testing phase of the mathematical optimization an elliptical negative exponential cost surface provided many characteristics of a generalized cost surface.

This same surface is used in this section to demonstrate the effect of cost on path location. The surface is defined by:
 $C(x,y)=K\exp(-Ax^2-By^2)$,(24)
where K,A, and B are positive constants. The behaviour of this surface is ideal for presenting problems within this section. At $C(x,y) = K$ at $x=y=0$. For positive A and B with large $|x|$ or large $|y|$; $C(x,y)$ approaches 0. The cost function $C(x,y)$ is defined everywhere in the plane implying that a path can locate anywhere without creating discontinuity problems.

4.1.1 Cost surface orientation

Figure 16 presents four paths, the end points of which are located along the x-axis. The major axis of the elliptical cost surface is located on the y-axis. The ability of a minimum cost path to avoid the higher cost areas is diminished as the end points approach the center of the cost surface. To avoid the high cost center, the path must initially traverse through the high cost region. In other words it sometimes is cheaper to traverse a high cost region than to avoid it. The cost versus distance plot in Figure 17 illustrates the cost profile of the path. In regions outside the high cost areas of the cost surface the paths are virtually straight lines due to the nearly homogeneous nature of the surface.

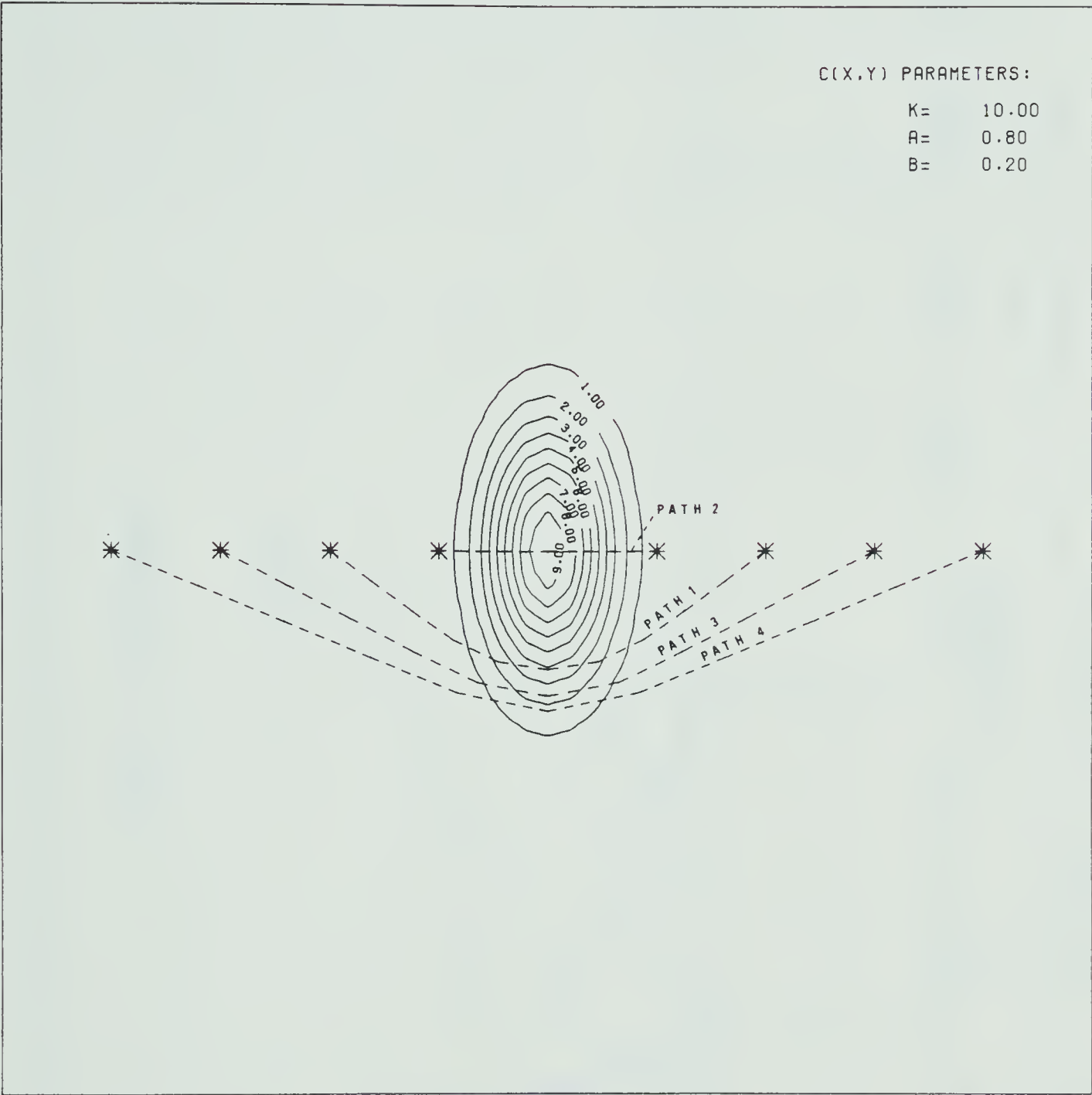


Figure 16 Four paths traversing the cost surface

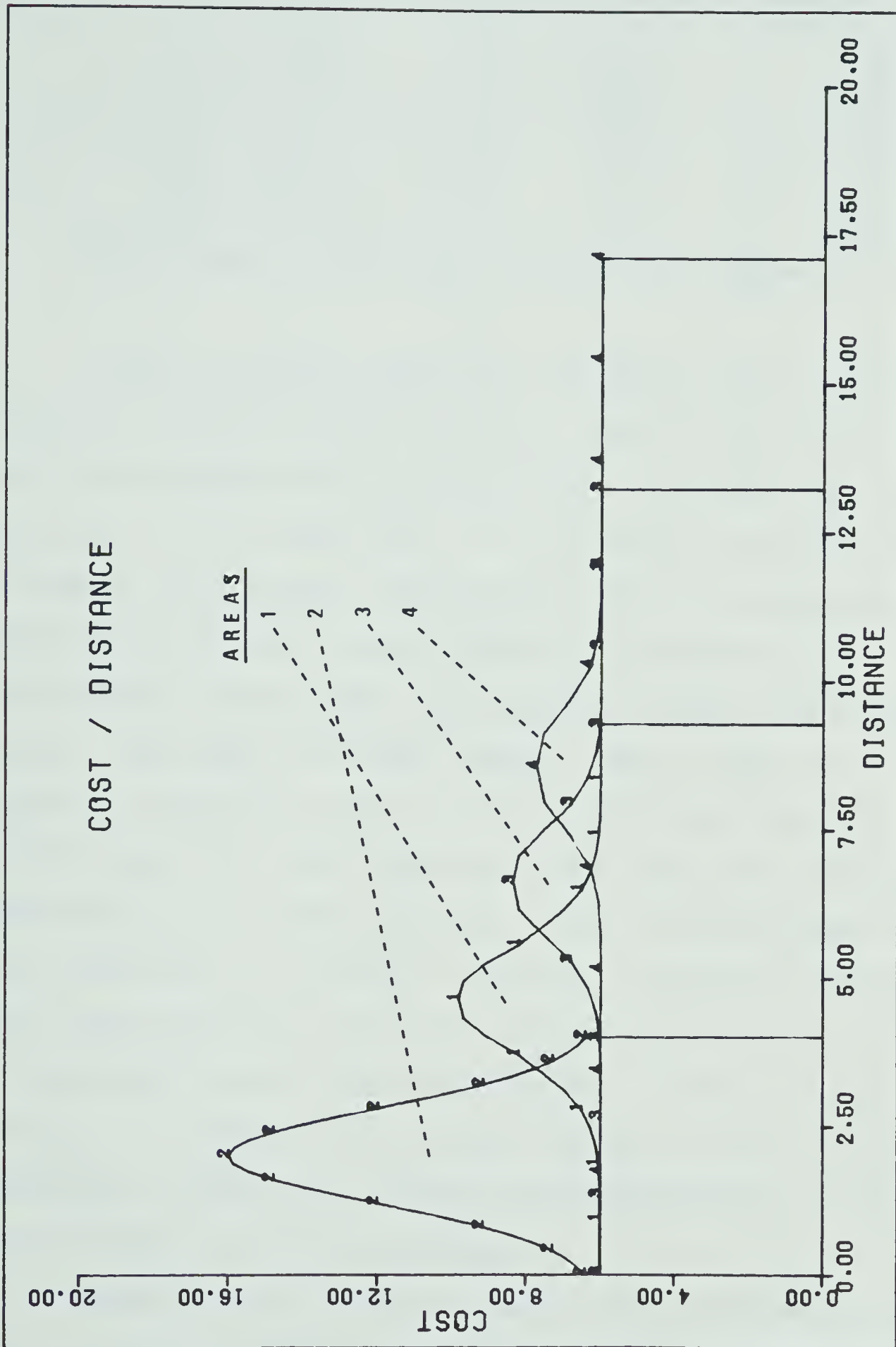


Figure 17 Profile of cost versus distance from one end point for paths in Figure 16

Path #	Base Cost	Cost	Distance	Cost/Distance	Sinuosity Ratio
2	6.0	43.59	4.00	10.90	1.00
1	6.0	64.55	9.28	6.95	1.16
3	6.0	84.95	13.23	6.42	1.10
4	6.0	106.74	17.11	6.24	1.07

Table 2 Cost and distance analysis for Figure 16

Table 2 provides numerical information for cost and distance variables concerning the paths in Figure 16. For each minimum cost path the base cost, the total cost (area), distance and the two ratios, cost divided by distance and sinuosity, are given. The sinuosity ratio is a measure of path curvature and is calclated by dividing the path distance by the straight line distance between the two end points. The ratio is always greater than or equal to 1.0. Table 2 illustrates that as the end points are located further away from the high cost center, the cost/distance and the sinuosity ratio decrease. Path 2, which cannot avoid the high cost of the center, chooses a virtually straight path resulting in a sinuosity ratio of 1.00. This illustrates the fact that the decision of whether to go through or to avoid the high cost regions of the surface is dependent on both the starting and ending points of the path and the shape of the cost surface.

The purpose of the next example is to illustrate the effect that the orientation of the cost surface has on path

location. The major axis of the surface is now centered on the x-axis. The end points remain the same as in Figure 16.

Path #	Base Cost	Cost	Distance	Cost/Distance	Sinuosity Ratio
2	6.0	47.46	4.00	11.86	1.00
3	6.0	36.95	6.01	6.15	1.50
1	6.0	42.11	9.59	4.39	1.20
4	6.0	53.83	13.07	4.12	1.09
5	6.0	68.14	16.80	4.06	1.05

Table 3 Cost and distance analysis for Figure 18

The effect of the orientation is illustrated by comparing paths in Figures 16 and 18. In both cases the cost/distance and the sinuosity ratio decrease as the end points locate away from the center. However, the relationship between the end points and the high cost regions result in different cost/distance profiles as demonstrated in Figures 17 and 19. The most illustrative profile is that for path 2 (Figure 16) shown in Figure 17 and path 3 (Figure 18) shown in Figure 19. Both paths have the same end points. Path 2 is located directly through the center. No effort is made to deviate to a lower cost region. The orientation of the cost surface suggests that such a deviation would be of little benefit, because the distance traveled to gain this region is too large and too costly. Path 3 in Figure 18, however, does deviate to a lower cost

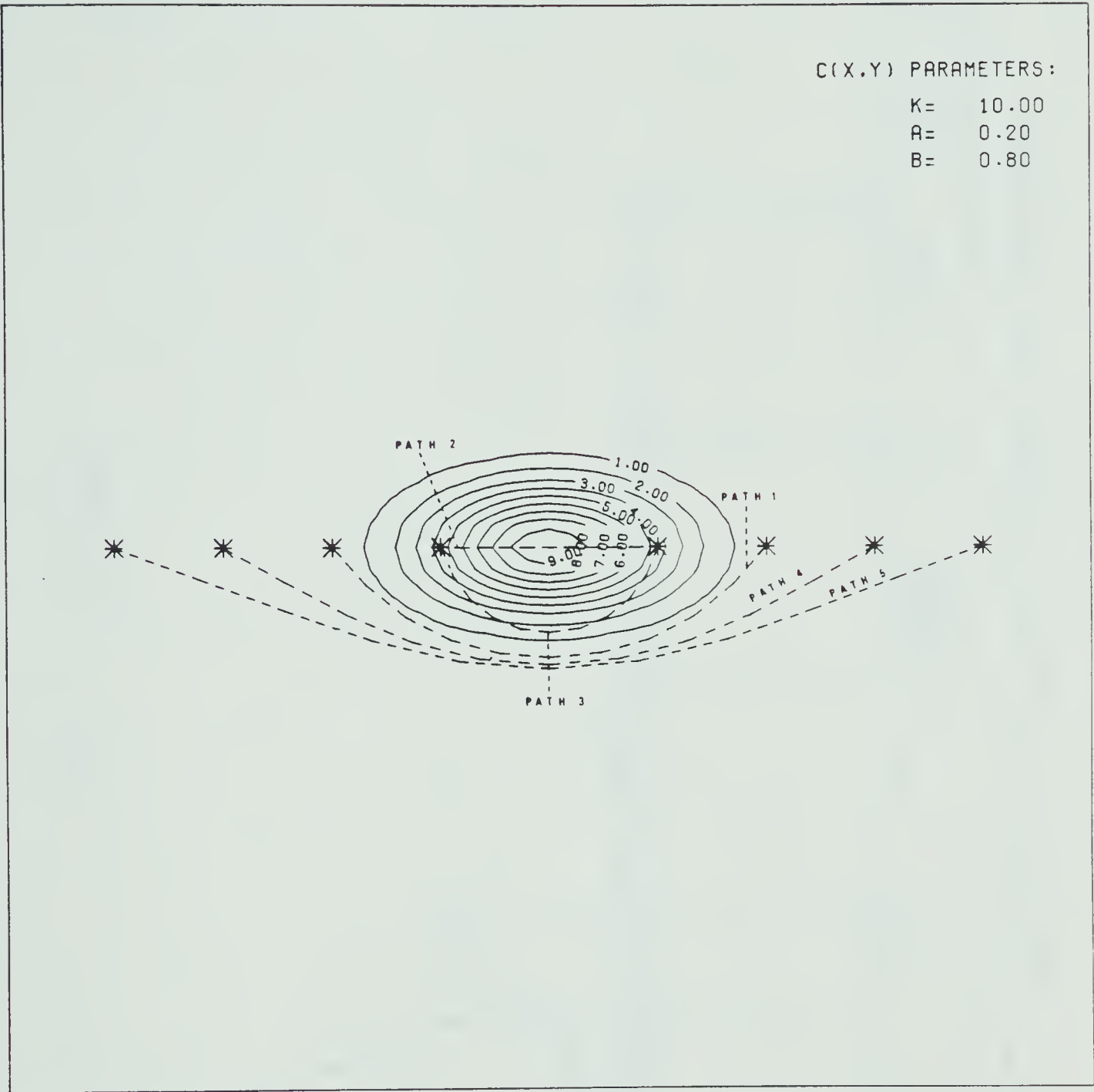


Figure 18 Four paths avoiding the high cost regions

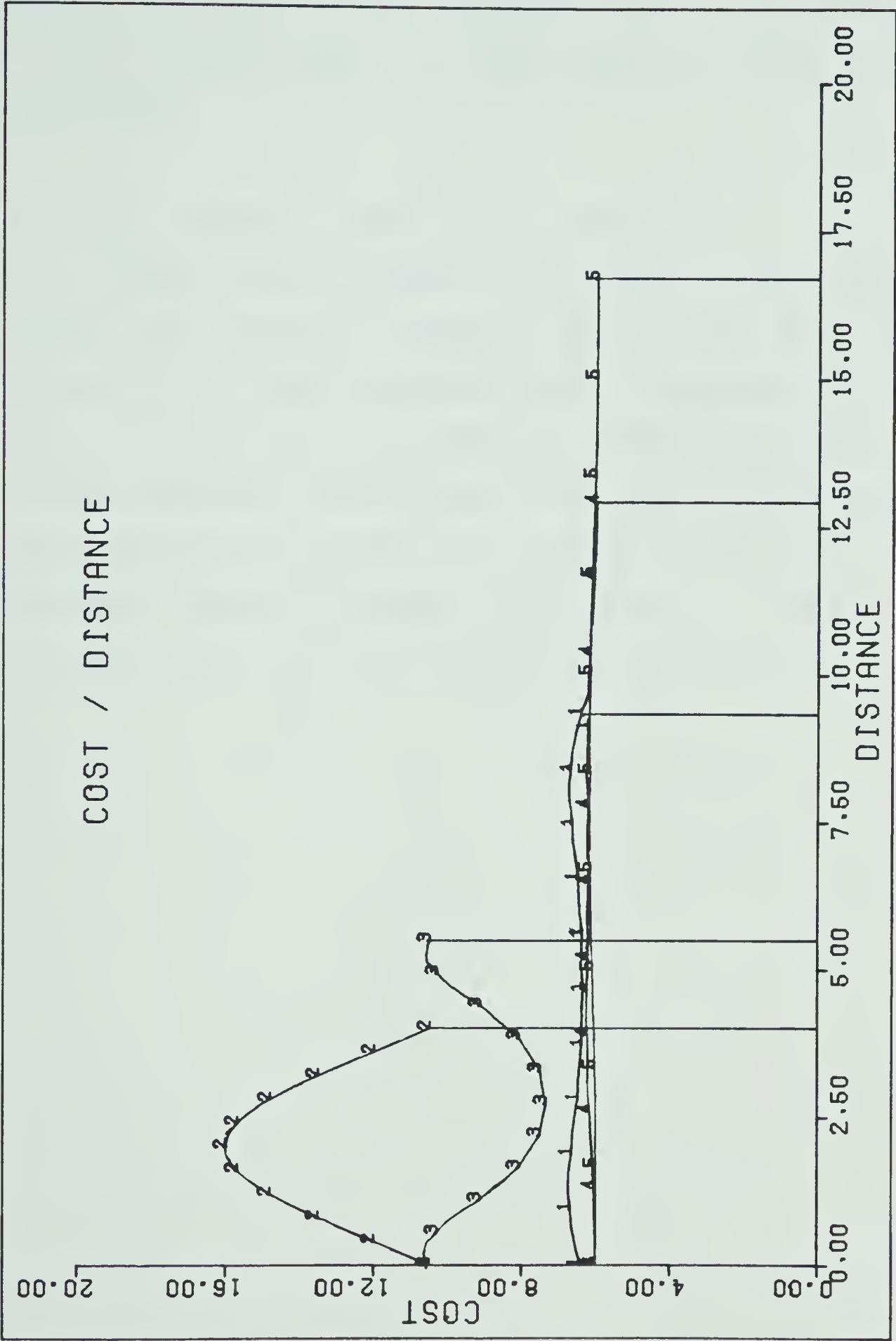


Figure 19 Profile of cost versus distance from one of the end point for paths in Figure 18

region and the extra distance of this path over the direct path 2 resulted in a cost saving of 10.51. The orientation of the surface allowed the higher deviation to be profitable.

4.1.2 The effect of cost on path curvature

The decision of whether to go around a high cost region of the cost surface or through it must be made based on the calculation of the minimum cost path. The purpose of the paths in Figure 20 is to demonstrate the effect of the cost on path curvature. The maximum curvature of the minimum cost path occurs in the region where avoiding high cost is still possible. The cost/distance profile given in Figure 21 indicates the relative areas occupied by each path.

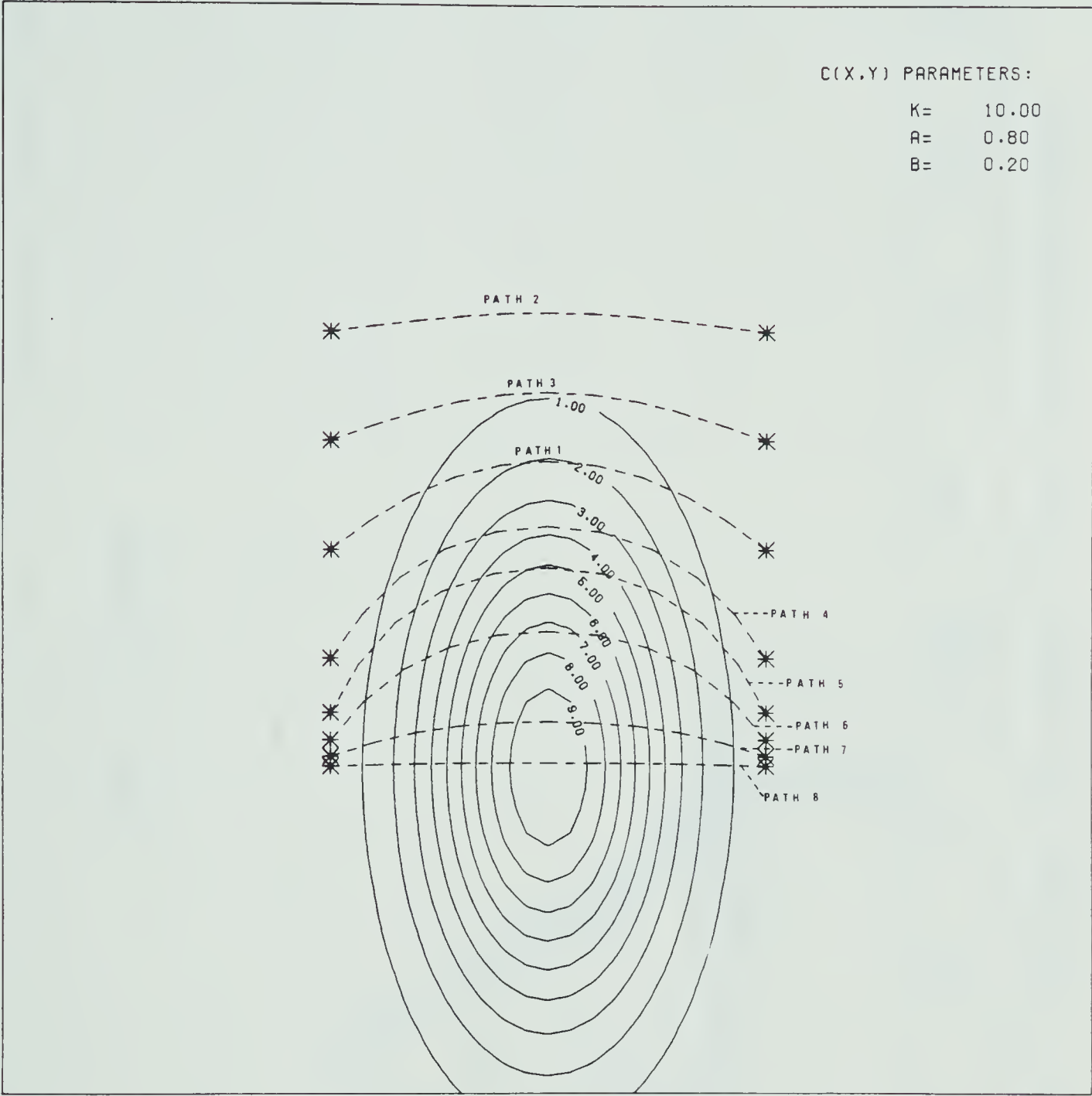


Figure 20 The effect of cost on path curvature

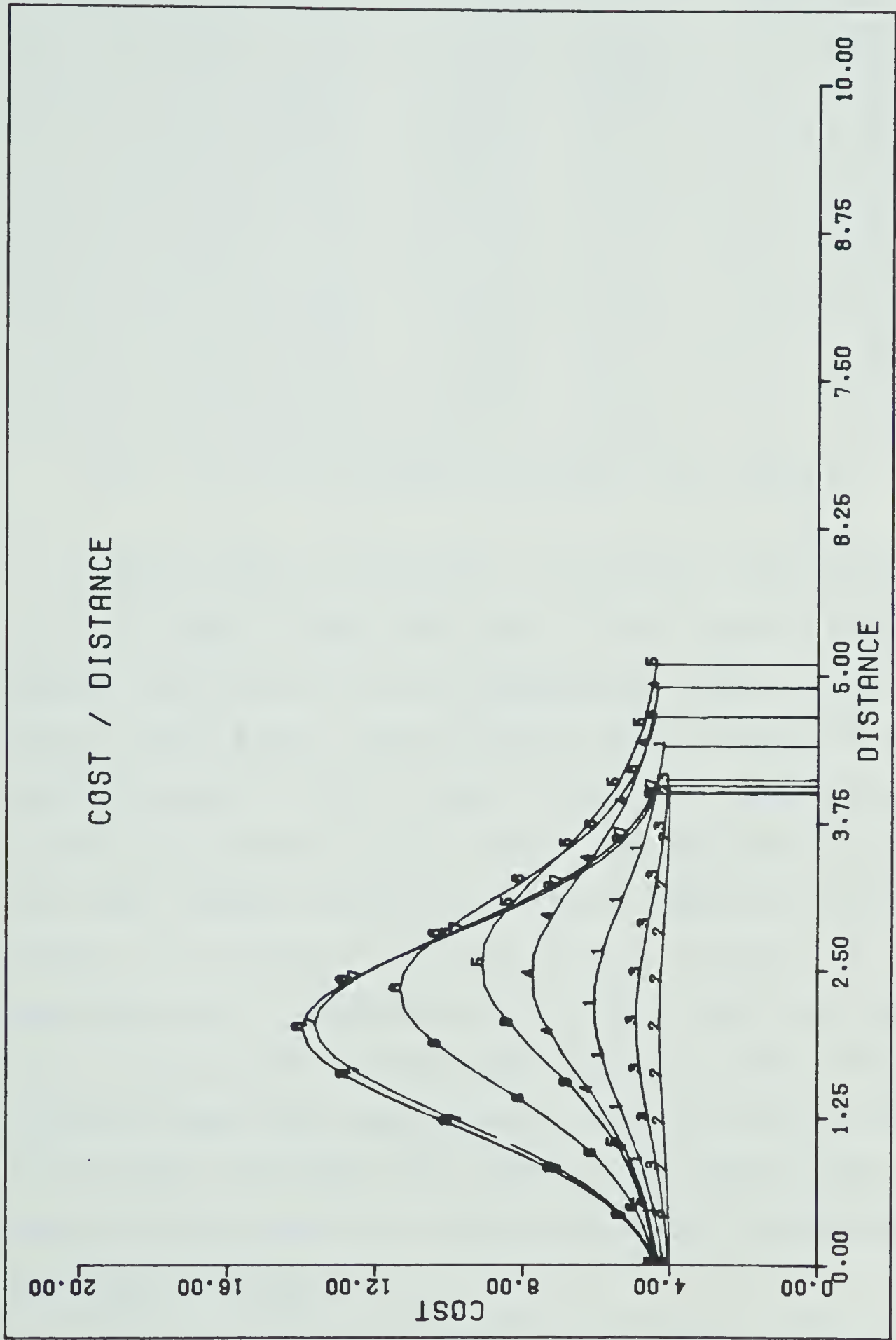


Figure 21 Profile of cost versus distance from one end point
for paths in Figure 20

Path #	Base Cost	Cost	Distance	Cost/Distance	Sinuosity Ratio
2	4.0	16.71	4.02	4.16	1.01
3	4.0	18.54	4.12	4.50	1.03
1	4.0	22.50	4.04	5.11	1.10
4	4.0	28.86	4.90	5.83	1.22
5	4.0	32.73	5.10	6.42	1.28
6	4.0	34.68	4.65	7.46	1.16
7	4.0	35.44	4.07	8.71	1.02
8	4.0	35.59	4.00	8.90	1.00

Table 4 Cost and distance analysis for Figure 20

The curvature of the paths is given by the sinuosity ratio in Table 4. Some paths have a low sinuosity ratio because they are in a nearly homogeneous region of the cost plane. Others have a lower sinuosity ratio because they cannot deviate to the low cost regions. This observation can be made from paths 2 and 8. These paths approach a straight line even though they are in different regions of the cost surface. The effect of the cost surface in path 2 is negligible as it contributes only 0.63 to the total cost. Path 8 cannot avoid the high cost region of the surface and chooses a path which deviates only 0.02 coordinate distance at $x=0$ from a straight line. Path 5, the longest path, has end points the position of which allows for the avoidance of the high cost areas.

4.1.3 Varying the base cost

One of the more interesting examples which can be used to illustrate the relationship between cost (area) and distance is provided by Figure 22. The base cost is increased progressively until it dominates the cost surface. The example verifies the obvious. A million dollar per mile divided highway is less effected by variations in local costs than a small country road which winds around every slough. The higher the base cost the straighter the minimum cost path, ceteris paribus.

Path #	Base Cost	Cost	Distance	Cost/Distance	Sinuosity Ratio
2	1.0	12.31	10.42	1.18	1.84
3	2.0	22.00	9.09	2.42	1.61
4	4.0	38.70	7.77	4.98	1.37
1	8.0	67.04	6.56	10.22	1.16
5	16.0	114.86	5.69	20.19	1.01

Table 5 Cost and distance analysis for Figure 22

The major observation for this example is shown by the sinuosity ratio. As the base cost increases, the sinuosity ratio decreases indicating that a route with higher base cost is less influenced by the cost surface.

4.2 Radially symmetric cost surfaces

The use of radially symmetric surfaces to represent

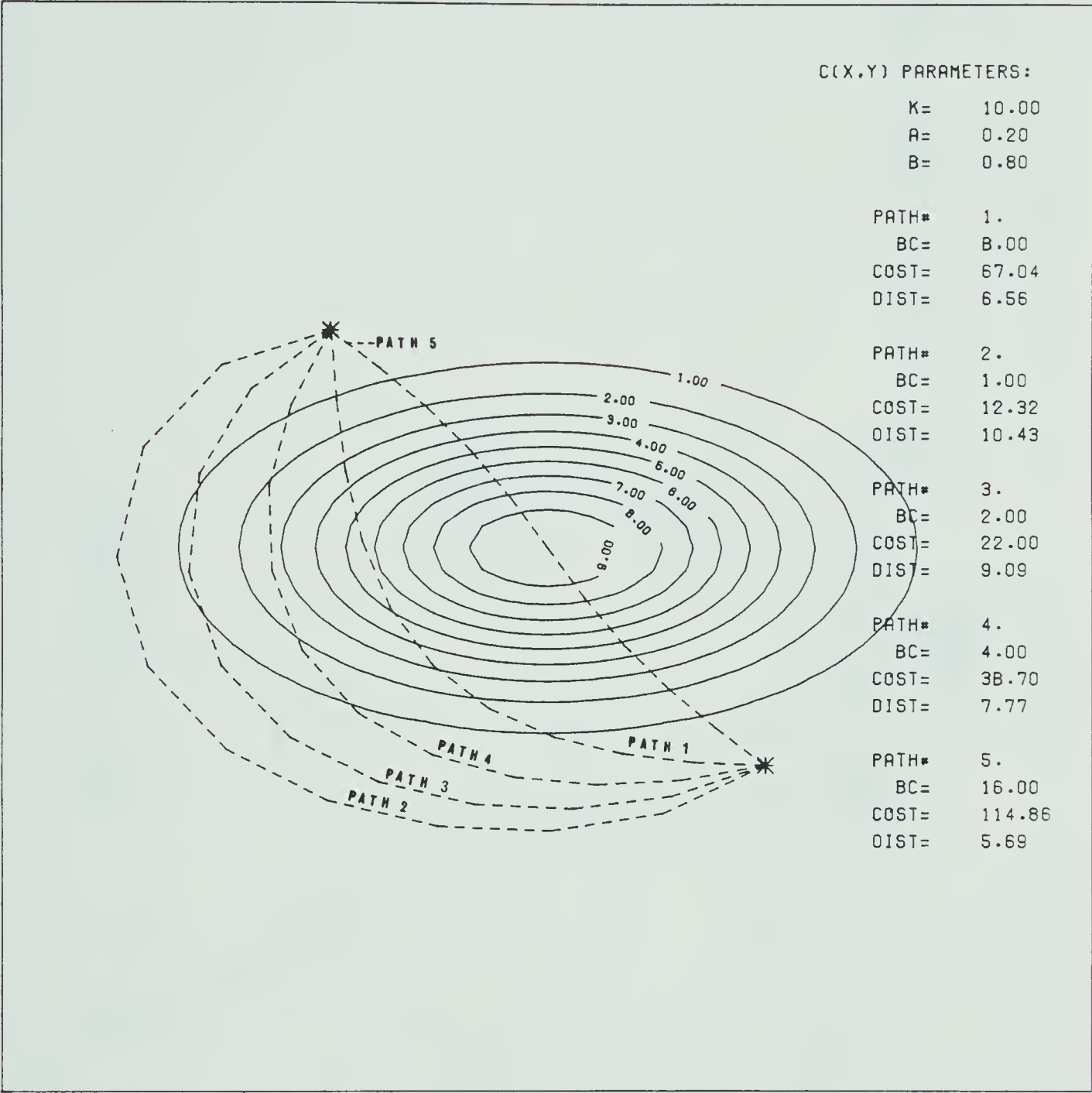


Figure 22 The increasing base cost (BC) example

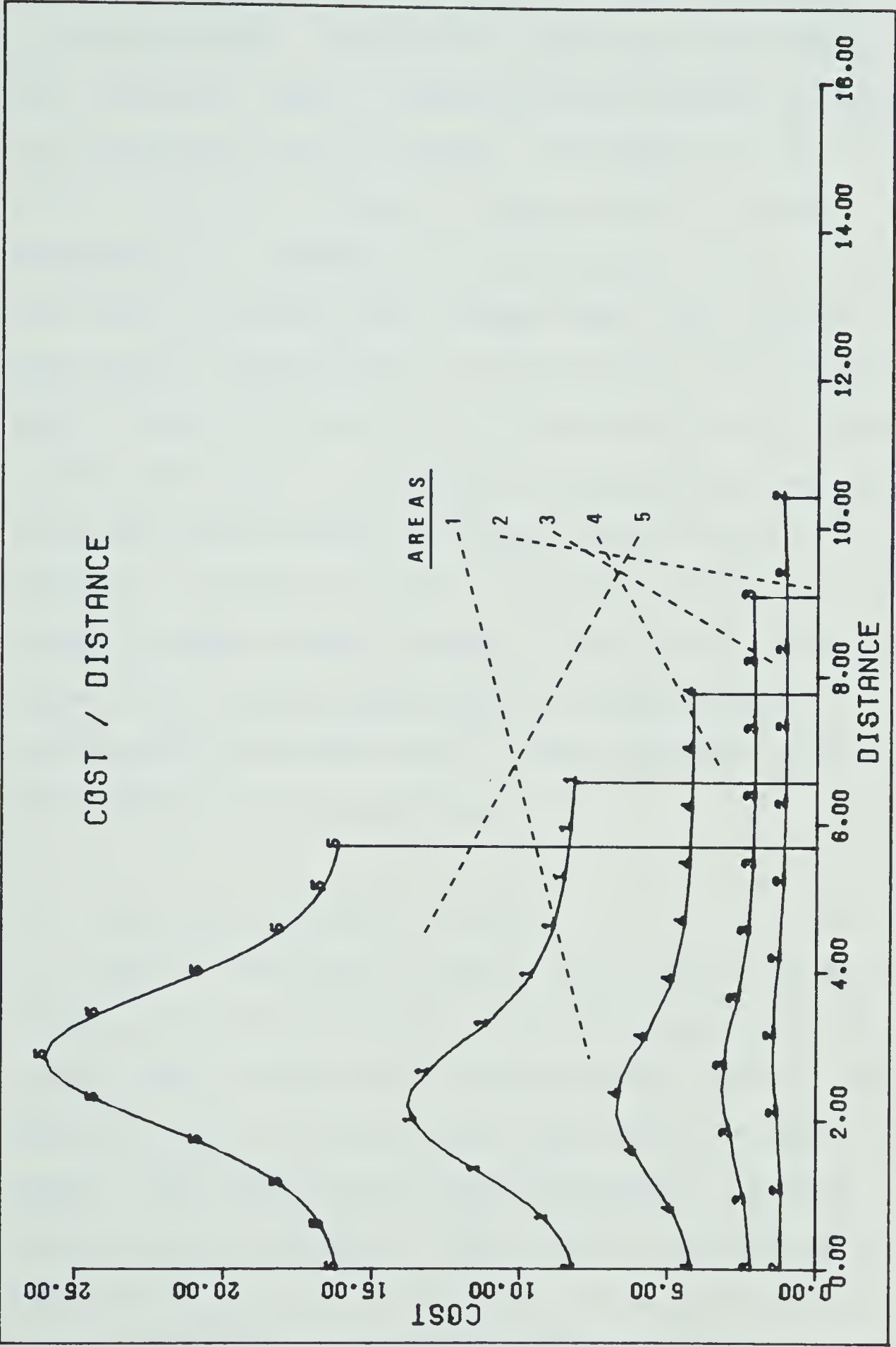


Figure 23 Profile of cost versus distance from one end point for paths in Figure 22

some cost variable within an urban setting has been given recent attention by Wardrop (1969), Angel and Hyman (1970,1972,1976), Zitron(1974), and Puu(1978a,1978b). The cost variable, usually measured as the inverse of velocity, is a function of the distance r from the city center. In order to provide a model of velocity that is both mathematically expedient and functional within the urban setting, the authors have assumed that the velocity is distributed symmetrically from the center. This assumption may not be valid for all cities, but as Angel and Hyman (1976) point out it is useful to illustrate a methodology which derives minimum cost paths. Presenting radially symmetric examples in this section has one purpose. The examples shown are well known in the current literature. By comparing the paths generated by the mathematical optimization to those found in the literature, the merits of both methods can be evaluated.

4.2.1 Wardrop's velocity surface

One of the radially symmetric surfaces used by Wardrop (1969) varies the velocity as the distance away from the center. The velocity is given by $V(r)=Kr$, where K is a constant and r is the distance from the city center. In this thesis, the inverse of velocity provides a suitable formulation of the cost (time) surface, which becomes:

$C(x,y)=1/Kr, \dots\dots\dots(25)$

where $r=\sqrt{x^2+y^2}$. In order to duplicate Wardrop's example, a

partial family of paths radiating from the point $P(1,0)$ is shown in Figure 24. This family is virtually identical to Wardrop's. Path 1 is in the form of a semi-circle from the point $P(-1,0)$ to $P(1,0)$. This path can be used as a further validation of the mathematical optimization method.

4.2.2 Angel and Hyman's velocity surface

The velocity surface used by Angel and Hyman (1976) was obtained from an empirical study of travel times by the SELNEC Transportation Study (1968) for the city of Manchester Great Britain. The velocity, expressed as a function of r , the distance from the city center, is given by:

$V(r) = A - B \cdot \exp(-K \cdot r), \dots\dots\dots(26)$

where A , B , and K are constants. For this example $A=24.9$, $B=16.9$ and $K=0.56$. The travel velocity at the city center is 8 miles per hour. The velocity approaches 24.9 as the distance from the center increases. The time surface, which is used to represent the cost surface is as in Wardrop's example, the inverse of the velocity.

Angel and Hyman (1976) generate a family of minimum cost paths radiating from a point, and in this thesis the same end point was choosen. The family of minimum cost paths (Figure 25) corresponds to the paths found in Angel and Hyman's work. All paths except the one from the center curve away from the high cost regions of the cost surface.

Apart from presenting their family of minimum cost



Figure 24 Minimum time paths for Wardrop's velocity surface

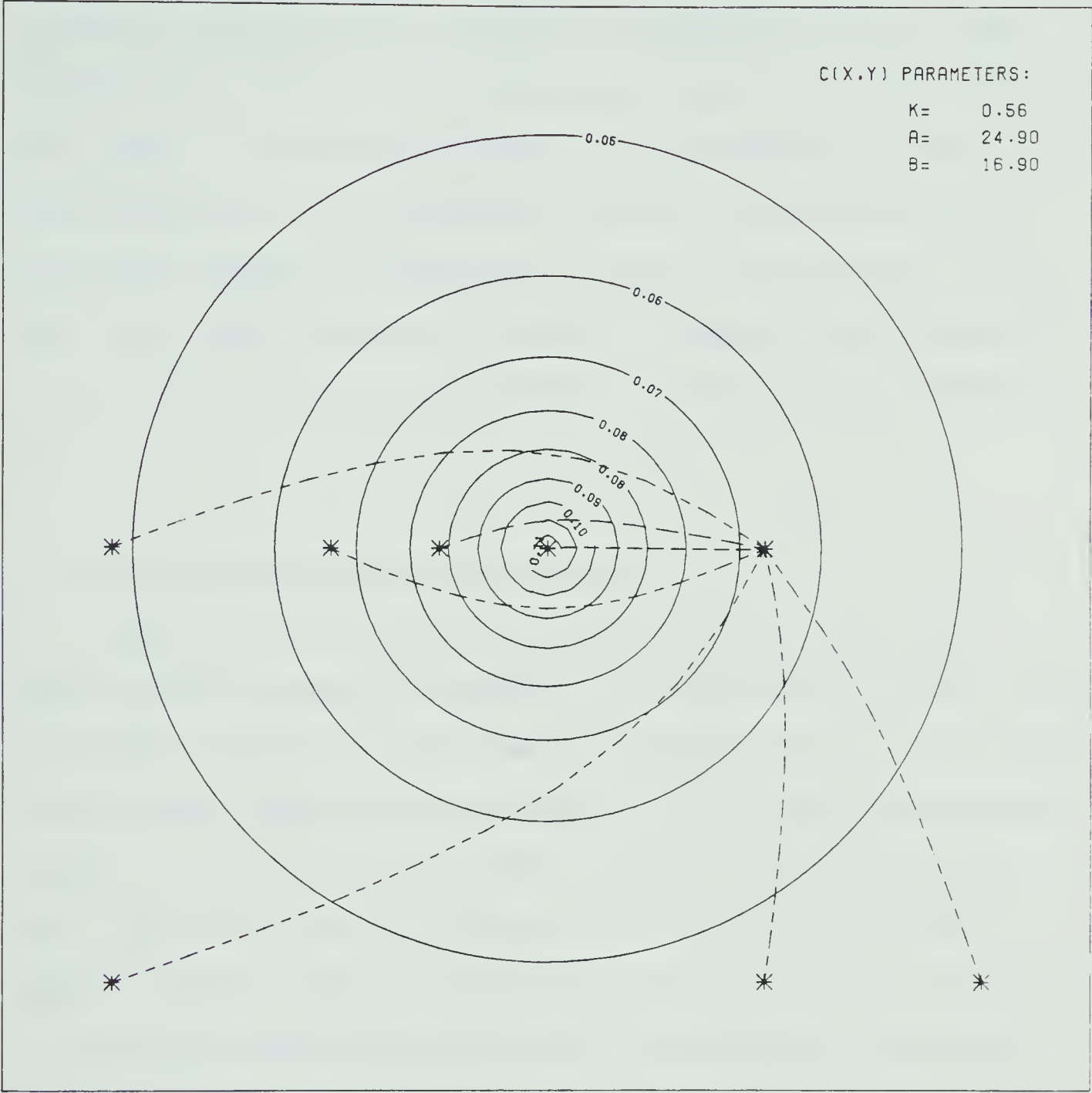


Figure 25 Minimum time paths for Angel and Hyman's velocity surface

paths, neither Wardrop (1969) and Angel and Hyman (1976) provide much discussion on the path locations. This thesis does not part with their tradition. The paths shown in Figures 24 and 25 are used solely to demonstrate that the mathematical optimization method can duplicate their minimum cost paths. This demonstration is significant, for the equivalence of the minimum cost paths illustrates that different methods can be used to solve the problem. In addition, these examples provide a framework for discussion of the transformation method which is discussed in chapter 5.

4.3 Constrained minimum cost paths

Some problems require that minimum cost paths be restricted to specific regions of the geographic plane. For this type of problem, the use of constraints to locate such minimum cost paths seems appropriate. This section presents examples of minimum cost paths constrained by linear and non-linear constraints. Minimum cost paths restricted to regions of the plane by straight line boundaries are derived by imposing linear constraints on the objective function. The barrier and corridor examples given below demonstrate linear constraints. Non-linear constraints, invoked by penalty functions, may be used for excluding paths from regions which can not be demarcated by straight lines. The minimum distance from a point constraint is provided as an example.

4.3.1 The barrier problem

The barrier problem is best presented by the following hypothetical case. Given a cost surface within two countries divided by an international boundary and the requirement that all paths originating and terminating in one country must remain in that country, then the minimum cost path is constrained by the border. The example in Figure 26 involve two countries, say Canada and the U.S.A., with a cost surface traversing the border given by $y=-1$. The path 1 constrained by the international border is 1.1 times as expensive as the minimum cost path 2. Although no cost surface is provided by the road map of British Columbia, the highway between Princeton and Cranbrook is a potential real world example of such a constraint in action. Highway #3 is forced to locate to the south of mountain ranges and lakes but is constrained to stay in Canada by the international border.

4.3.2 Minimum cost paths contained within a corridor

The location of transportation facilities within a corridor is common in both urban and rural settings. Constraints on the dimensions of the corridor may restrict the optimum location of minimum cost paths. The example in Figure 27 illustrates a corridor defined within two straight lines at $y=+1$ and $y=-1$. As in the previous example linear constraints may be applied to the objective function such

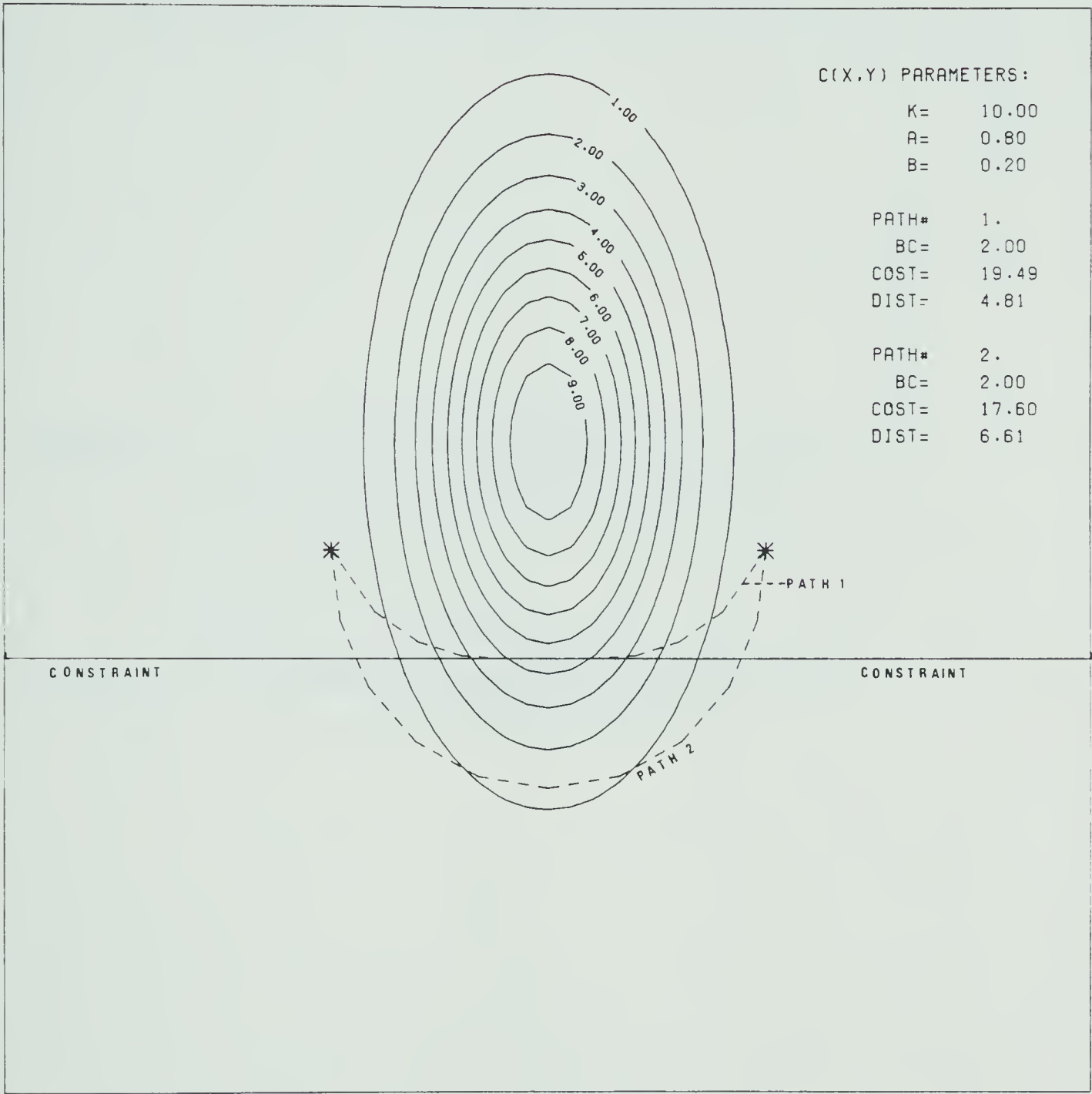


Figure 26 Path constrained by a barrier

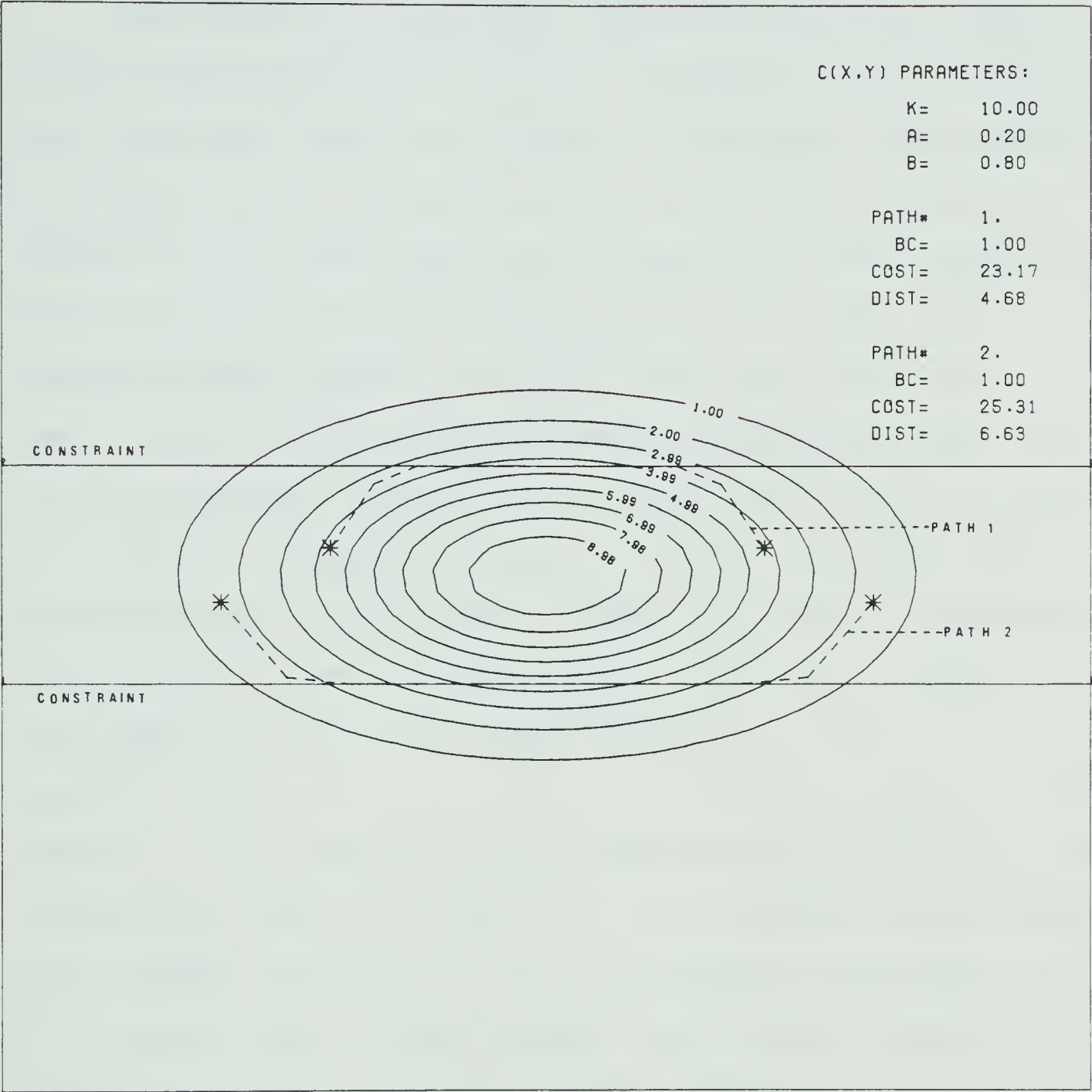


Figure 27 Two paths contained within a corridor

that the optimum path will locate within the corridor.

4.3.3 Fixed distance penalty functions

Some types of constraints cannot be expressed linearly and may be incorporated into the objective function through the addition of penalty functions. In practice these penalty functions are extreme additions to cost within certain regions of the plane and a path traversing these regions would be subjected to virtually an infinite cost. The example chosen assumes two points and a path with the restriction that the path must never be located closer than a fixed distance (radius) from either point. This situation could arise if a pipeline was not allowed to locate within a given distance of an Eskimo village. The penalty functions are applied in the following manner. The cost is added to the objective function by the penalty:

$$\exp(25.0(r-d)), \dots\dots\dots(27)$$

where r is the radius and d is the distance of the path from a particular point (village). If $(r-d)$ is positive the added cost becomes large and if $(r-d)$ is negative the added cost is virtually zero. Hence the path would locate outside the region defined by the point and the radius.

Figure 28 contains a cost surface with two constraints imposed. Paths 1 and 2 are calculated with a base cost (BC) of 4.0 and Paths 3, 4, and 5 are calculated with a base cost of 8. Only Path 5 is unconstrained. The location of each path is dependent on the starting solution. If the starting

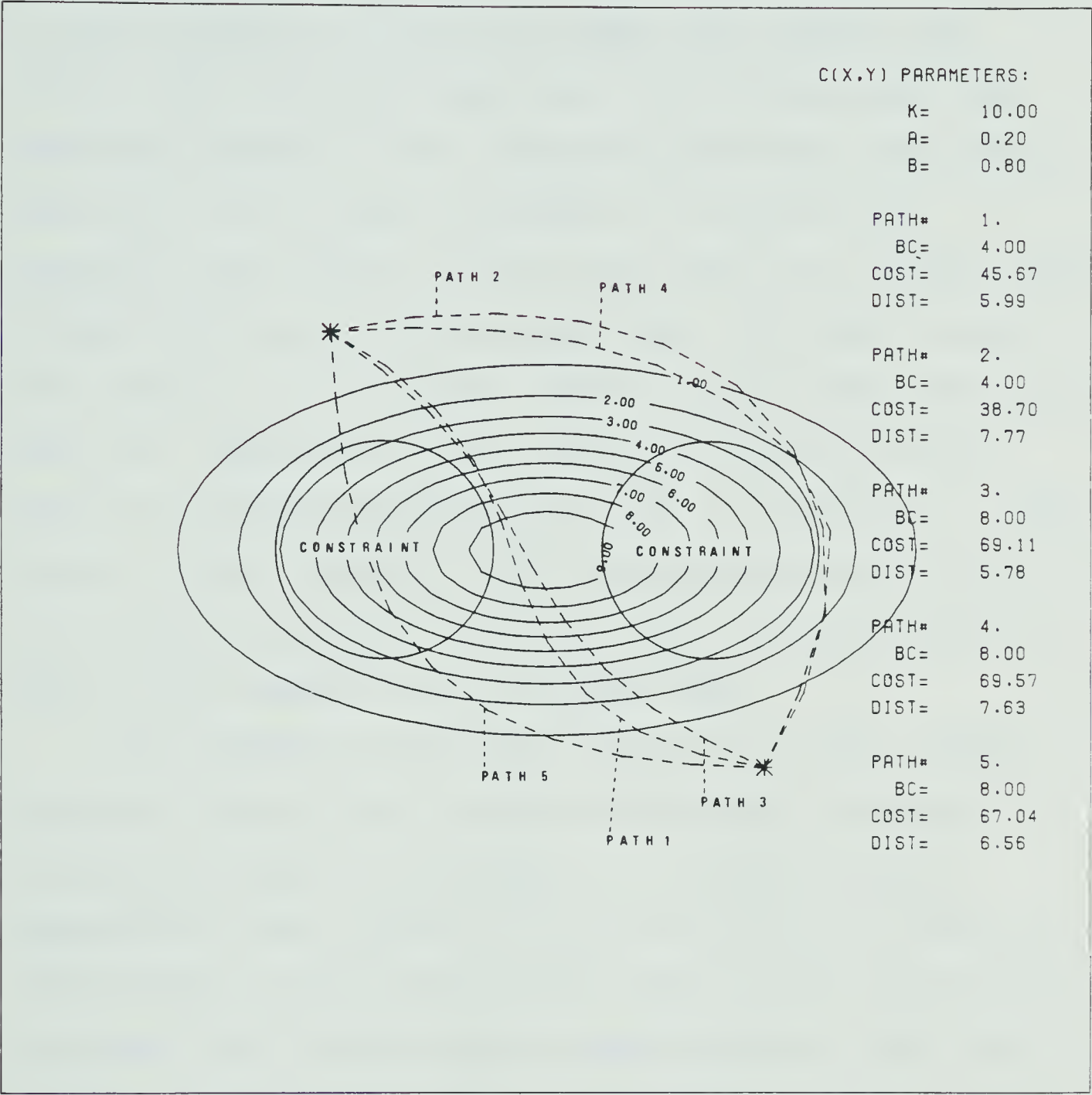


Figure 28 Paths constrained by circular regions

solution is located between the constraints, then a local minimum was found in the same region. Similarly, if the starting solution is located on one side of the constraint, a local minimum is found on the same side. Before the optimal path can be declared, both local minima must be found and compared. Path 1 located between the constraints cost 45.67, while Path 2 located to the top of one constraint cost 38.70. In this case Path 2 is a better local minimum and therefore the optimal path. For Paths 3 and 4 the situation is in reverse. Path 3, between the constraints cost 69.11 and Path 4 to the top of the constraint cost 69.57. The unconstrained optimal path is located as in Path 5 and a similar symmetric path exists between Paths 3 and 4. This unconstrained path indicates that the constraints do restrict the location of the optimal path.

The location of a minimum cost path for this example cannot be left entirely to the optimization. As shown the optimization routines converge on a variety of paths depending on the location of the starting solution and all combinations of paths through the constrained regions must be tested. Only three feasible combinations of paths exist in this example; around each of the constraints and between them. In figure 28 the placement of the constraints with respect to the end points and the cost surface is symmetric, therefore, only two combinations of paths required testing. Other problems may require the use of both combinatorial and optimization techniques for a solution. These types of

problems may be topics for future research.

$$S = \int_{P_0}^{P_f} ds \quad , \quad \dots\dots\dots (28)$$

where $ds = \sqrt{(dx)^2 + (dy)^2 + (dz)^2}$. This problem also requires a calculus of variations approach and the solution to the appropriate Euler equations for most surfaces cannot be found analytically. For general cases, some numerical technique must be used to evaluate the integral and the optimization techniques employed in this thesis are appropriate.

The integral (28) may be formulated as:

$$\text{minimize } F = \sum_{i=0}^{n-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2} \quad , \quad \dots\dots (29)$$

where n equals the number of links in the discrete path representing s. The objective function is optimized using the variable r method and is represented by:

$$F(T, r) = \sum_{i=0}^{n-1} \sqrt{r^2 + (z_{i+1} - z_i)^2} + w(\sqrt{(x_f - x_{n-1})^2 + (y_f - y_{n-1})^2} - r)^2 \quad , \quad \dots\dots\dots (30)$$

where $z_{i+1} = z(x_{i+1}, y_{i+1})$ and $z_i = z(x_i, y_i)$. The representation of T, r, x, y, and w is defined in Section 3.2.3. Because the multiplier effect of the cost term is absent from the objective function, its optimization is less difficult than calculating the minimum cost path.

5.1.2 Geodesic examples

Great circles calculated between two points on a sphere are geodesics. Because of their important classical role in cartography, the first example illustrates a geodesic on a hemisphere. The second example demonstrates the geodesic on a cone.

The great circle curve from Edmonton to London is shown in Figure 29. The contours of the surface correspond to the latitude. The projection of both the latitudes and the geodesic on the equatorial plane is analagous to an orthographic projection. (It was not derived by conventional means, using the transformation equations.) Both Edmonton and London are connected to the pole via meridians which are also great circles. According to the Atlas of Alberta, the actual distance between Edmonton and London is 6,796 kilometers. The geodesic distance calculated by the optimization routines is 6,825 kilometers. This 29 kilometer difference requires an explanation.

A number of errors resulted in the formulation of the Edmonton to London great circle. The x and y coordinates which are located on the equatorial plane were rounded to the nearest hundred kilometers. The second problem concerned representing the radius of the earth. The earth is a geoid which bulges at the equator. The radius calculated as the distance from the equatorial plane to the pole is smaller

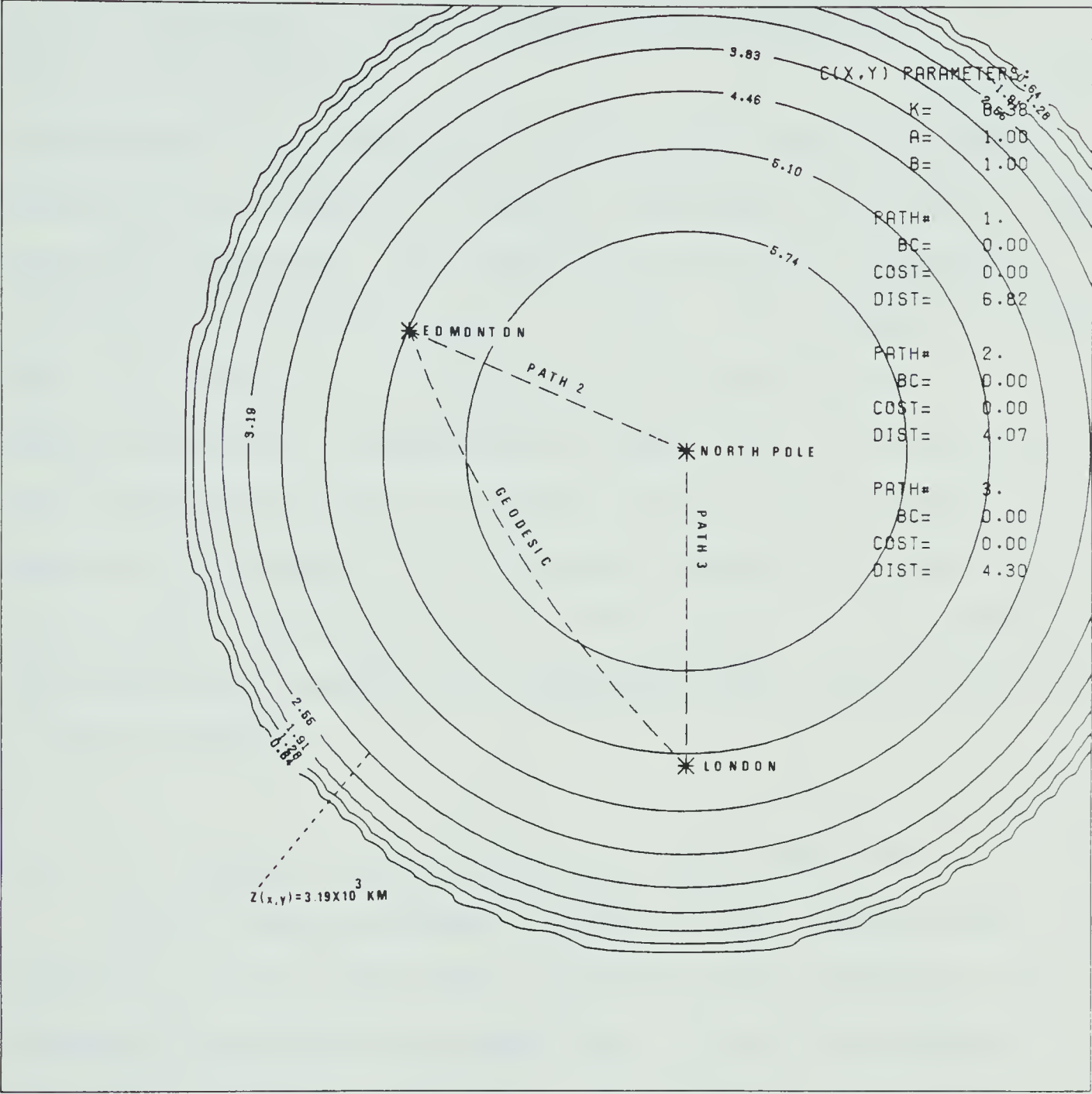


Figure 29 The Edmonton - London geodesic

than that on the equatorial plane. This smaller radius would result in a shorter great geoid distance.

The geodesic shown in Figure 30 is calculated from $P(-1,-1)$ to $P(+1,+1)$ for a conic surface given by:
$$Z(x,y)=1.5 \sqrt{x^2 + y^2} \dots\dots\dots(31)$$
 Fortunately, the geodesics for this cone can be expressed analytically. Appendix I contains the analytical solution to the Euler equations. The theoretical distance calculated analytically in this example is 3.901. The distance of 3.899 was calculated by the optimization routines, which demonstrates their preciseness. Another way of calculating the distance would be to develop the cone on a plane and measure the straight line distance (geodesic) between the two points. The eigenvalues of the Hessian for the objective function are all positive, indicating that this geodesic is a local minimum.

5.1.3 The geodesic and minimum cost path compared

During the preliminary stages of this research the minimum cost path was often compared to the geodesic. Two reasons invalidate this comparison. First, the geodesic is measured in terms of distance, the minimum cost path in terms of area. Secondly, the geodesic is located on a surface whereas the minimum cost path is located on the x,y plane. A demonstration of the location of the paths relative to the surface contours can be made by projecting the geodesic onto the x,y plane. Figure 31 presents the minimum

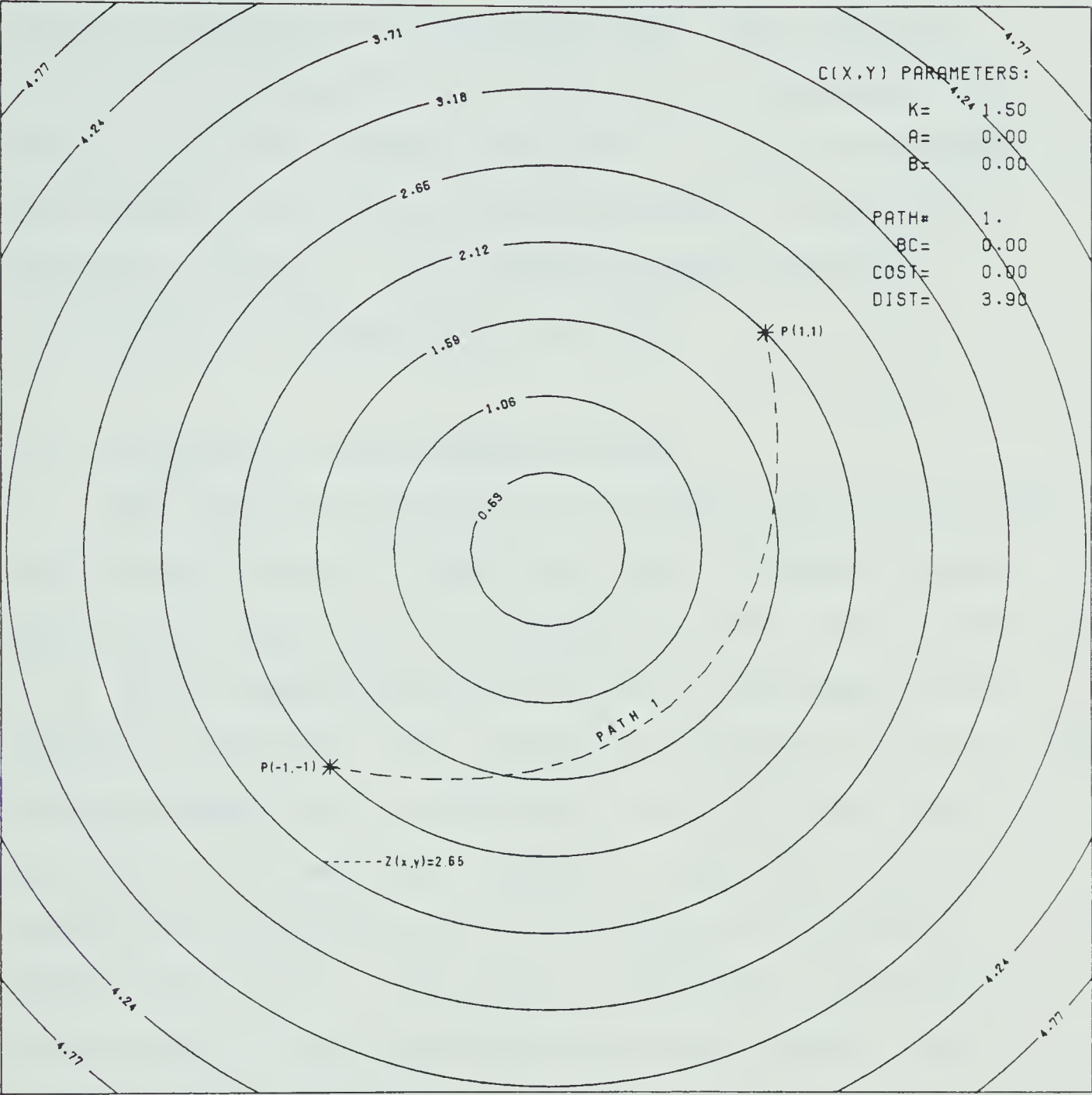


Figure 30 Geodesic on a cone

cost path (path 1) and the geodesic (path 2). For this particular example the two are projected differently onto the x,y plane. In addition, the geodesic is about three times as long as the minimum cost path. For radially symmetric surfaces with one end point at the center the projection of the geodesic onto the x,y plane corresponds to the minimum cost path as can be observed in Figure 32. Generally, however, the projected geodesic does not correspond with the minimum cost path.

5.1.4 Geodesics as minimum cost paths

Some types of minimum path problems can be formulated to minimize distance rather than cost. In order to give geographical meaning to the geodesic within this context the x , y , and z coordinates must all be of the same physical dimension and scale. This condition restricts the use of the distance measure as a substitute for cost to applications for which a 3-dimensional surface is defined; such as calculating the great circle distance on the globe as demonstrated above. This class of problem is significant from a purely theoretical point of view, however, and warrants some exploration.

A theoretical example of this problem is presented in the following scenario. Assume a sinuous landscape in either the x or the y direction. The problem is to determine the location of a minimum length path between two points on the surface. Examples of such problems can be found in the

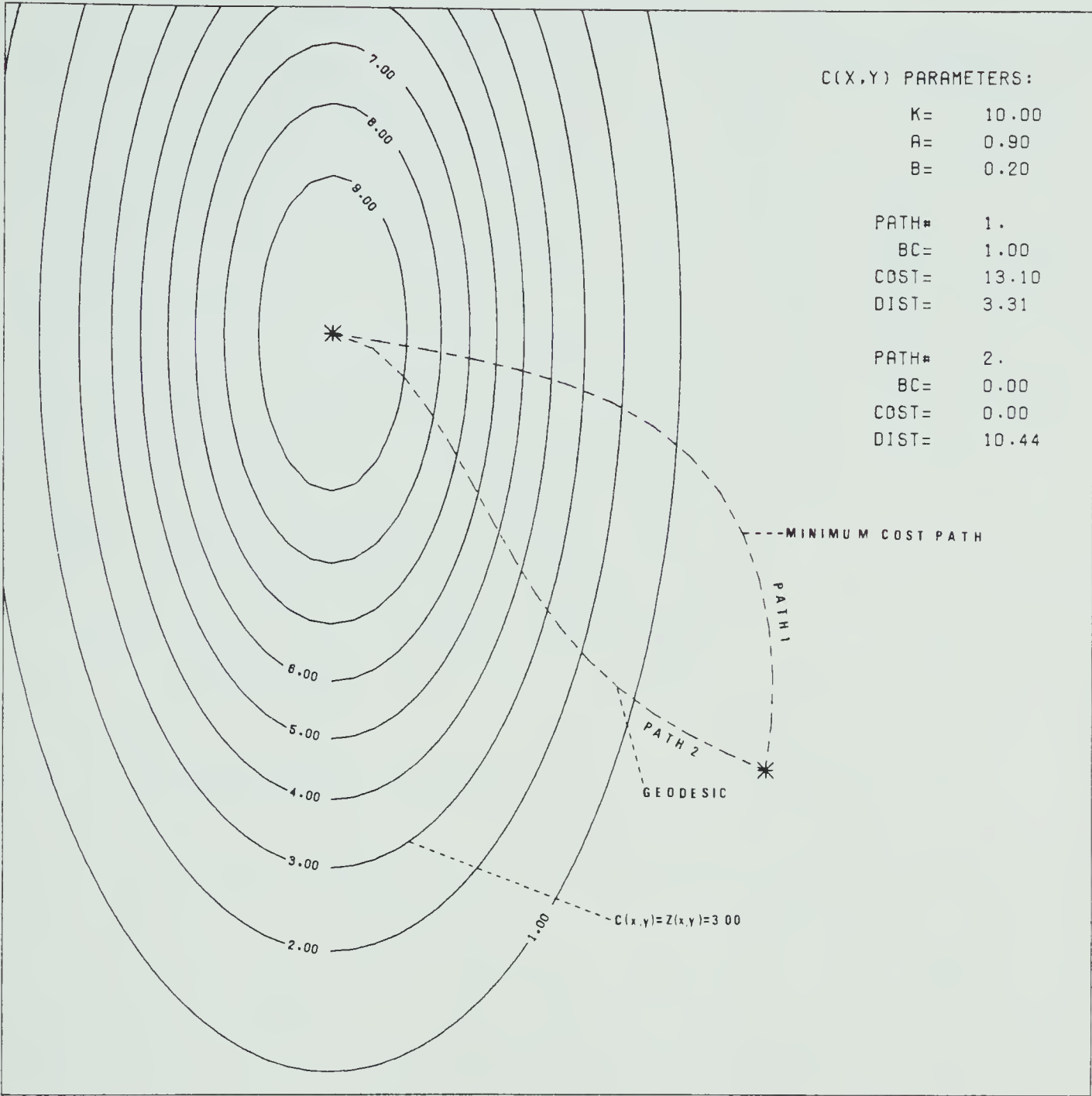


Figure 31 Geodesic and the minimum cost path compared

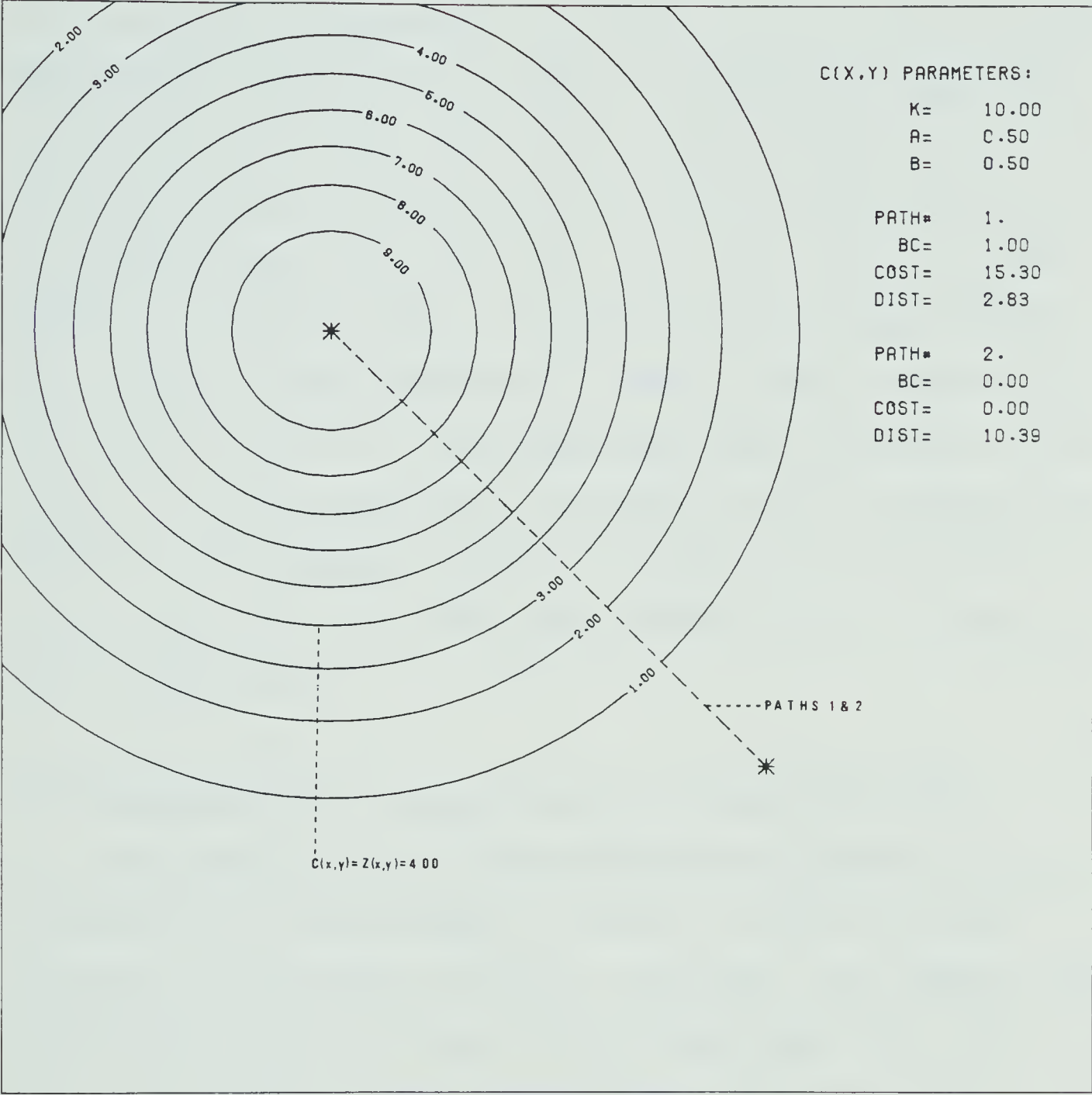


Figure 32 Geodesic and the minimum cost path compared

process of locating new pipelines or utility corridors.

Figure 33 demonstrates three geodesics for a surface given by:

$Z(x,y)=4\sin(0.1x)\cos(y)+4.(32)$

The geodesic represented by path 1 is calculated from the point $P(x,y,z)=(-6,6,1.83)$ to $(6,-4,2.52)$. Path 2 is calculated from $(-8,-5,3.19)$ to $(8,5,4.81)$ and path 3 is calculated from $(-8,-6,1.24)$ to $(-8,6,1.24)$. Each of these geodesics seems to locate in regions of the surface where the change in the z coordinate is small. This is especially noticable in path 3. A path generated above a straight line on the x,y plane generates a distance of 26.16, whereas the length of the geodesic is 22.10. A result such as this illustrates that it is sometimes shorter to walk around a hill, then over it.

5.2 Transformations, geodesics and minimum cost paths

The previous section demonstrated the geodesic as an alternative to the minimum cost path. For some problems, such as finding great circles on the globe, the geodesic may be the only feasible form of the minimum cost path. The natural extension to this observation is to find the conditions for which the geodesic can be substituted for the minimum cost path. The observations in the previous section imply that the geodesic which represents length must also represent cost. The surface, on which the geodesic is calculated, must therefore equate cost and distance. The

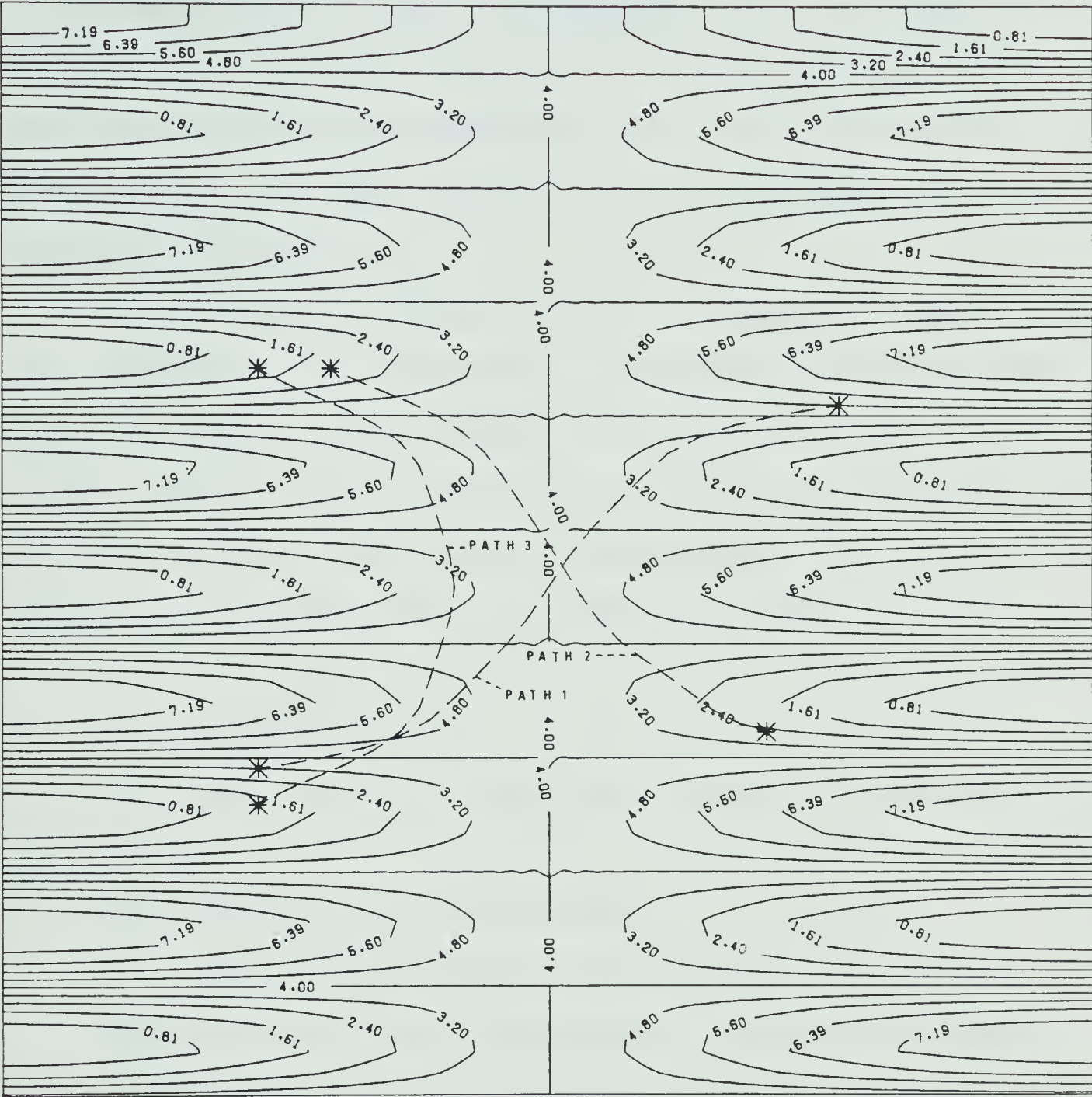


Figure 33 Three geodesics on a sinuous surface

type of cost surface considered by this thesis does not meet this condition. The z values of this surface indicate cost at some coordinate location x and y . If this cost surface is to be used to derive the cost geodesic, all x, y , and z coordinates must depict locations in the cost space. This implies that the cost surface must be transformed to a three-dimensional cost space if it is to be used for calculating geodesics.

The purpose of this section is to summarize and discuss the results of the transformation research of Warntz (1968), Angel and Hyman (1972a, 1972b, 1976), and Puu (1977, 1978a, 1978b). Their research addresses the problem of transforming the cost surface into another surface where upon the geodesic corresponds to the minimum cost path.

5.2.1 Puu's contribution

The theme of the transformation method is presented by Puu (1978b):

The concept of a cost surface

The optimal paths in general turn out to be curved. Owing to this the natural question has been posed whether it is possible to map the original region, R , onto some other region S in such a way that the curved paths are mapped onto straight lines in some sense and so that distance in the image region equals cost.

Wardrop (1969) has investigated the

possibilities of mapping R onto some other plane region S by conformal mappings, that is, by complex analytic functions with nonvanishing derivatives. Angel and Hyman (1970, 1972, 1976) on the other hand, guided by conjecture by Warntz (1967), explore the possibilities of mapping R onto a curved surface S , embedded in three-dimensional euclidean space, where the images of the optimal paths are 'straight' in the sense of being geodesic. Their discussion too is confined to conformal maps, characterized by the facts that angles are preserved and that magnification is independent of direction.

The reasons why discussions are confined to conformal maps are never stated, but I am going to demonstrate that, in fact, conformal mapping is the only type that works with isotropic transportation problems. Angel and Hyman, moreover, only discuss the special case of surfaces of revolution. They suggest that the method has a much greater degree of generality than Wardrop's, but the exact field of application never becomes clear. In fact the general assumptions they use turn out to be too restrictive.

Puu then continues with theorems and conclusions to demonstrate three concepts concerning the transformation method.

Puu concludes that "the cost for any isotropic

transportation problem is a conformal transformation of the plane". An isotropic transportation problem is a model in which the cost of transportation or movement is defined at a point. The cost surfaces explored by Warntz, Wardrop, Angel and Hyman, and in this thesis are all isotropic transportation models. Puu's transformed cost surface is the conformal map of these isotropic models. Next Puu develops the necessary conditions which must be met by the cost surfaces such that the geodesic on the transformed cost surface corresponds to the minimum cost path. He verifies that the surfaces used by Angel and Hyman meet the necessary conditions for a transformation and duplicates their examples.

Puu's final conclusion states that it is "always possible to find a (local) cost surface such that distances equal cost in an isotropic transportation problem". He observes that from these local effects nothing can be concluded about the global nature of the cost surfaces which generally can have many folds and complex curvatures.

5.2.2 The three steps in the transformation process

The actual calculation of minimum cost paths using the transformation methodology is a three step process. The first step requires the formulation of the transformation which maps the original cost surface into one in which distance equals cost. The second step requires the calculation of the geodesic on the transformed surface. The

third step uses the reverse transformation to map the geodesic into the minimum cost path on the original geographic plane. These three steps warrant some discussion because using this process Angel and Hyman, and Puu find minimum cost paths for only two cost surfaces.

The conformal transformations for a specific cost surface where cost is expressed as a function of x and y is in general not easily achieved. Even if the surface has this isotropic characteristic, the transformation requires the application of Gaussian differential geometry. According to Tobler (1961) this is a very difficult topic and best avoided. These difficulties are evident in the work of Angel and Hyman (1972 b), and Puu (1978 b). Both authors have transformed the same surfaces. The radially symmetric velocity surface of Wardrop (1969), which is demonstrated in Section 4.2.1, becomes a cylinder when transformed. The velocity surface $V(r)=Ar^2 + B$, where A and B are constants, and r is the distance from the city center is transformed into the sphere. Although Puu (1978b) provides the necessary conditions which must be met by the cost surface before it can be transformed from the region R to the region S , his examples stop at the radially symmetric cost surfaces of Angel and Hyman. He shows that these surfaces also meet the necessary conditions for the transformation, but indicates that the transformations themselves are difficult to achieve.

The second step in finding the minimum cost path

requires that the geodesic on the transformed surface be found. The methods illustrated in Section 5.1 of this thesis can be used once an analytic form of the surface is given. Angel and Hyman, and Puu provide analytic surfaces for which the geodesic are well known. The cylinder produces helices and the sphere produces great circles. Generally however, the geodesics are not easily found by intuitive or analytic methods. Angel and Hyman (1970) use Huygens (1912) method for calculating geodesics. Isochrones, lines of equal value from a point, are constructed using the original cost surface. The family of geodesics from this point is normal to each isochrone. Angle and Hyman (1972) state that Huygen's construction requires considerable adaptation before it can be a practical method for computing minimum cost paths for general surfaces.

The third step concerns mapping the geodesic back into the original geographic plane from the cost space such that a minimum cost path can be plotted. This mapping depends on the initial transformation: if it exists, then an inverse transformation should be found which can map the geodesic into the minimum cost path. No attempt has yet been made to use this inverse mapping method.

The difficulties in applying the three step transformation method is illustrated by Angel and Hyman's (1972 b) velocity surface. Although their surface is radially symmetric the transformation could not be achieved analytically and required numerical methods. Once the

transformed surface was derived, the geodesics could not be found and the transformation was used only to illustrate the form of the time surface. Angel and Hyman used Huygen's construction to plot the isochrones for this velocity surface. Minimum cost paths as shown in Figure 25 were drawn perpendicular to these isochrones.

5.2.3 Results of the transformation methods

One result of Angel and Hyman's (1972 b) transformation process is the proof that the minimum cost path of the original surface corresponds to the geodesic on the transformed surface. This mathematical proof clearly links the minimum cost path to that of the geodesic. There are theoretical advantages in providing this link. The intuitive conjecture of Warntz (1965) was shown to be true. "A transformation exists which will map a realistic distribution of transportation costs on the Euclidean plane into a curved surface with uniform transport facility" (Angel and Hyman (1972 b)). The validation of this conjecture provided new impetus to the transformation methods. Warntz's example of minimum land acquisition routes in the United States is used as reference in much geographic transportation and path research. According to Harvey (1969) this research is important because the geographic concept of distance can no longer be viewed as the two dimensional Euclidean distance, but must be extended to more dimensions to take into account the activities which may effect this

distance. The transformation methods contribute to this exploration of the distance theory in geography.

There are however major difficulties in pursuing the transformation methods for deriving minimum cost paths. The difficulty lies in the mathematical conditions which must be met by the cost surface. Although Puu (1978 b) provides the necessary conditions, he provides no example beyond those of Wardrop (1969), and Angel and Hyman (1972 b). Also, no example of transforming a non-symmetric cost surface is found. Puu indicates that transformation can always take place locally for an isotropic cost surface, but he suggests that providing the actual transformations and geodesic is very difficult. Thus, if non-radially symmetric cost surface such as the ones used in Chapter 4 are to be transformed, major mathematical research must be attempted.

Chapter 6 Conclusion

This thesis solves two problems in the calculus of variations; the minimum cost path problem and the problem of finding the geodesic on a given surface. For each of these problems an objective function is formulated and analyzed, and a number of case studies are performed. This chapter summarizes the results of using mathematical optimization techniques to solve these problems.

6.1 The methodology

Perhaps the most significant contribution of this thesis to minimum path research is found in the methodology of deriving the minimum cost path. Although the representation of the cost integral by an objective function is not new, the development of the r and θ method and the use of constraints have not yet been explored in the literature. This section discusses the merits and disadvantages of using the mathematical optimization method to solve the Euler-Lagrange equations.

6.1.1 The method works

Representing the cost distance and the distance integral by an objective function which is then minimized is a technique that works. Two analyses referred to in this thesis support this conclusion. The trajectory method of Rankin (1979) calculated a similar minimum path to that

calculated with the variable r and θ method. This comparison is presented in section 3.2.5. The second analysis is shown in section 5.1.2. The length of the geodesic is calculated using both analytic and optimization methods. The results differed by a distance of 0.002. Thus two independent methods were used to verify the representation of the integrals (3) and (28) by the objective function (23) and (30) respectively.

Confidence in the variable r and θ objective function provides many benefits for minimum cost path research. Many different path problems can be posed and solved as demonstrated in Chapter 4. The method is flexible and can deal with the interaction of the end points, cost surface and base cost. The varying base cost problem shown in section 4.1.3 is an example of this interaction. The current path literature has shown families of minimum paths for radially symmetric surfaces, but has not presented problems in which the base cost interacts with the cost surface. In addition, the method verifies the obvious minimum cost paths, such as the straight line of the homogeneous cost surface, and the geodesic lines of longitude from Edmonton to the north pole and from London to the pole.

Minimizing a non-linear objective function with many variables also presents some problems regarding the acceptance of a minimum solution. Although there are strong indications that the objective functions are convex, the

minimization of some specific problems resulted in solutions which could not be accepted as optimum. For example, the cost surface represented by $C(x,y)=K\exp(-Ax^2-By^2)$ becomes virtually zero for large $|x|$ or $|y|$. The derivatives of $C(x,y)$ with respect to x or y in this region are also near zero. If no base cost is added to this function a path has virtually zero cost and can locate anywhere in the plane. A zero gradient is one of the necessary conditions for accepting a minimum cost path. Hauer's (1974) algorithm indicated a zero gradient for solutions with large $|x|$ or $|y|$ but simple intuitive reasoning indicated that such solutions are meaningless in reality. The solution methods could not differentiate between a zero gradient and a zero derivative of the cost surface. The Hessian, however, for this case is not positive definite and no local minimum is claimed. The addition of a base-cost resulted in more reasonable minimum paths.

6.1.2 The constraints

The derivation of the minimum path, which is subjected to some type of locational constraint is sometimes very difficult. Current literature avoids the discussion of this topic entirely and the examples presented in this thesis were of a very simple exploratory nature. The reason for the difficulty lies more in the practical than in the theoretical nature of constraints. It is easy to draw a line over which the minimum path cannot cross, but it is more

difficult to formulate this constraint into the solution procedures of the optimization. These difficulties are overcome for two types of constraints. The simple linear inequality constraint and the representation of a constrained region by penalty functions. The examples presented in section 4.3 illustrate initial success at the use of constraints.

The locational constraints expressed as linear inequality constraints presented the fewest problems to Hauer's (1974) algorithms. Provided that the initial starting solutions were within the feasible or unconstrained region, the solution was found easily. Difficulty, however, was encountered in applying these linear constraints to the variable r and θ objective function. The decision variables for this method are not easily subjected to constraints imposed on the location of the path. Instead the variable x, y objective function formulation is used

Some locational constraints could not be expressed by linear constraints and penalty function were applied. Section 4.3.3 presented an example in which two circular regions of the plane were subjected to a cost penalty. The minimum paths were located outside this region indicating that the penalty function worked. There is however a topological problem related to this problem. The minimum path is found by an iterative procedure which depends upon the starting solution. A starting solution between the two penalized regions resulted in a minimum path also located in

the same region. Similarly, if the starting solution is located on one or the other side of the penalized region, then the minimum path is located within the same region. Conceivably, if the number of such penalized regions is large, the number of solution paths which must be tested is dependent on the number of possible feasible paths through these regions. This is a combinatorial problem and each feasible path must be derived from the optimization. However, given a particular starting solution and the use of a penalty function, the minimum path reflected the topology of the starting solution. The acceptance of this solution as the global optimum one can not be made. Only if all possible solution topologies are tested could such a decision be reached.

6.2 The transformation method

The limited analysis and discussion of the transformation methods given in this thesis is justified for two reasons. One, the methods are difficult in terms of the explanatory mathematics required and further development of these methods is out of the scope of this thesis. Two, the ability of the transformation method to solve the variety of problems posed in Chapter 4 of this thesis has not been shown, indicating that it is still in a development stage. The mention of these methods in Chapter 5 is however relevant, because the calculation of the geodesic is a topic that is discussed in the current transformation literature.

6.3 Future research

6.3.1 The cost surface

This thesis represents cost by a continuous function of x and y . Angel and Hyman (1972 b) provide a rough fit for their negative exponential surface and refer to the work of Clark (1951) for further support. It is recognized that the real cost space is in general very complex and any surface representing this space may contain many irregularities. In order to satisfy the continuity condition required by the optimization procedures only limited options are open for representing realistic cost surfaces.

A piecewise continuous cost surface representing local regions may possibly be worked into the solution procedures. A redesign of the decision variables in the objective function may be in order for this approach. Finally a three dimensional representation of cost represented by $C(x,y,z)$ is explored by Werner and Boukidis (1963) and Steenbrink (1974). The type of cost surface is very realistic for mountainous environments. Again the decision variables of the objective function must be modified or expanded to include the third dimension.

6.3.2 The use of constraints

Two immediate projects concerning constraints can result from this thesis. The first project could consist of

expanding the use of linear constraints beyond those presented in Sections 4.3.1 and 4.3.2. The second project could investigate the general use of penalty functions to represent a wider range of non-linear constraints than is given in Section 4.3.3. The topological problem discussed earlier may be part of this project.

6.3.3 Other path problems

Chapter 4 and 5 demonstrated the feasibility of using objective functions to find minimum cost paths and geodesics. By modifying these functions a variety of other path related problems may be solved. Typically, these problems may deal with the minimization or maximization of some interaction between a path and a surface. The attention given to such path problems in current geographic literature indicates that both the problems and their solution methodologies require further attention. The use of the mathematical optimization methodology to solve these problems is as shown in this thesis both realistic and feasible.

Bibliography

- Angel, S., and Hyman, G.M., 1970, "Urban velocity fields", Environment and Planning, 2, 211-224.
- Angel, S., and Hyman, G.M., 1972a, "Transformations and geographic theory", Geographical Analysis, 4(4), 350-367.
- Angel, S., and Hyman, G.M., 1972b, "Urban spatial interaction", Environment and Planning, 4, 99-118.
- Angel, S., and Hyman, G.M., 1976, Urban Fields, Pion, London, England.
- Aoki, M., 1971, Introduction to Optimization Techniques, Fundamentals and Applications of Nonlinear Programming, Macmillan, New York.
- Barber, G.M., 1970, "Variational calculus approaches to spatial design problems", in Seminar on Quantitative Geography, edited by Mackinnon and Scott, a University of Toronto Department of Geography Discussion Paper Series No.7.
- Burghardt, A.F., 1969, "The origin and development of the road network of the Niagara Peninsula, Ontario, 1770-1851", Annals of the Association of American Geographers, 59(3), 417-440.
- Bussiere, R., Snickars, F., 1970, "Derivation of the negative exponential model by an entropy maximising method", Environment and Planning A, 2, 295-301.
- Chorley, R.J., and Haggett, P., 1967, Models in Geography, Methuen and Company Ltd., London, England.
- Clark, C., 1951, "Urban Population Densities", Journal of the Royal Statistical Society, 114, 490-496.
- Courant, R., and Robbins, H., 1941, What is Mathematics?, Oxford University Press, London, New York, Toronto.
- Dantzig, G.B., 1966, All shortest routes in a graph, Technical Report, 66-3, Operations Research House, Stanford University.
- Dijkstra, E.W., 1959, "A note on two problems in connexion with graphs", Numerische Mathematik, 1, 269-271.
- Farbey, B., Land, A.H., and Murchland, J.D., 1967, "The cascade algorithm for finding all shortest distances in a directed graph", Management Science, 14, (September).

- Floyd, R.W., 1962, "Algorithm 97, shortest path", Communications of the Association of Computing Machinery, 5, 345.
- Goodchild, M.F., 1977, An evaluation of lattice solutions to the problem of corridor location, Environment and Planning A, 9, 727-738.
- Harvey, D., 1969, Explanation in Geography, St. Martin's Press, New York.
- Hauer, J.F., and Descheneau, J.C., 1973, MORPAK-I: A Mathematical Optimization Package for Unconstrained Minimization, Department of Computing Science, University of Alberta, Technical Report TR73-9.
- Hauer, J.F., 1974, MORPAK-II: An Accelerated Projection Program for Constrained Nonlinear Minimization, Department of Computing Science, University of Alberta, Technical Report TR74-13.
- Hettner, A., 1952, "Verkehrsgeographie", Allgemeine Geographie des Menschen, iii, 32, (Stuttgart).
- Howard, B.E., Bramnick, Z., and Shaw, J.F.B., 1968, "Optimum curvature principle in highway routing", Journal of the Highway Division, Proceeding of the American Society of Civil Engineers, 94, HW 1(June).
- Kuhn, H.W. and Tucker, A.W., 1951, "Nonlinear Programming", Proceedings of the 2nd Berkely Symposium, (Berkely, Calif.: University of California Press), 132-38.
- Losch, A., 1954, The Economics of Location, New Haven: Yale University Press.
- Losch, A., 1962, Die Raumlische Ordnung der Wirtschaft, Gustav Fischer Verlag, Stuttgart.
- Lowe, J.C., and Moryadas, S. , 1975, The Geography of Movement, Houghton Mifflin Company, Boston.
- Moore, E.F., 1959, "The shortest path through a maze", Proceedings of an International Symposium on the Theory of Switching, Part II, APRIL 2-5, 1957, The Annals of the Computation Laboratory of Harvard University 30, Harvard University Press, Cambridge Mass.
- Murchland, J.D., 1967, "The 'once through' method of finding all shortest distances in a directed graph and for the inverse problem", Transport Network Theory Unit Report LBS-TNT 56, Graduate School of Business

Studies, London, England.

Niedercorn, J.H., 1971, "A negative exponential model of urban land use densities and its implications for metropolitan development", Journal of Regional Science, 11(3), 317-326.

OECD., 1973, Optimization of Road Alignment by Computer, OECD Road Research Group, Paris.

Planning and Transportation Research and Computation, 1969, Proceedings of the PTRC-Symposium: 'Cost Models and Optimization in Road Location, Design and Construction', London, England (25-27 June).

Planning and Transportation Research and Computation, 1971, Proceedings of the PTRC-Symposium: 'Cost Models and Optimization in Road Location, Design and Construction', London, England (8-11 June).

Puu, T., 1977, "A proposed definition of traffic flow in continuous transportation models", Environment and Planning A, 2, 559-567.

Puu, T., 1978a, "On traffic equilibrium, congestion tolls, and the allocation of transportation capacity in a congested urban area", Environment and Planning A, 10, 29-36.

Puu, T., 1978b, "On the existence of optimal paths and cost surfaces in Isotropic continuous transformation Models", Environment and Planning A, 10, 1121-1130.

Puu, T., 1978c, "Towards a theory of optimal roads", Regional Science and Urban Economics, 8, 203-226.

Rankin, J., 1979, Personal Correspondence.

Ratzel, F., 1912, "Anthropogeographie", Die Geographische Verbreitung des Menschen, II, 347, (Stuttgart).

Ross, D., and Schemakcer, B., 1965, "Integrated Civil Engineering System (ICES)", Proceedings of the 1965 Conference on Improved Highway Engineering Productivity, May 18-21, Boston, Mass.

Schureman, L.R., 1965, "The Total Integrated Engineering System (ICES)", Proceedings of the 1965 Conference on Improved Highway Engineering Productivity, May 18-21, Boston, Mass.

Smith, D.M., 1966, "A theoretical framework for geographical studies of industrial location", Economic Geography,

42, 95-113.

Steenbrink, P.A., 1974, Optimization of Transportation Networks, John Wiley and Sons, London.

SELNEC, 1968, "Study area characteristics", Technical Working Paper 4, South East Lancashire and North East Cheshire Transportation Study, Manchester, England.

Tobler, W., 1961, Map Transformations of Geographical Space, Unpublished Ph.D. Thesis, Department of Geography, The University of Washington.

Tobler, W., 1963, "Geographic area and map projection", Geographical Review, 53, 59-78.

Turner, A.K., and Miles, R.D., 1971, "The GCARS System: a computer-assisted method of regional route location", Highway Research Record, 348, 11-15.

Turner, A.K., 1971, "GCARS: An approach to computer aided route selection", Proceedings of the Planning and Transport Research and Computation Symposium: 'Cost Models and Optimization in Road Location and Design', London, England(June).

Wardrop, J.G., 1969, "Minimum cost paths in urban areas", Strassenbau und Strassenverkehrstechnik, 86, 184-190.

Warntz, W., 1965, "A note on surfaces and paths and applications to geographical problems", Discussion Paper 6, Michigan Inter-University Community of Mathematical Geographers, Ann Arbor, Michigan.

Warntz, W., 1967, "Global science and the tyranny of space", Papers and Proceedings of the Regional Science Association, 19, 7-19

Weber, A., 1909, Theory of the Location of Industries, Chicago: University of Chicago Press.

Werner, C., and Boukidis, N.A., 1963, Determination of Optimum Route Connecting Two Locations, U.S. Army Transportation Research Command, Contract No.DA-44-177-Tc-685.(Transportation Center, Northwestern Univ.).

Werner, C., 1968, "The law of refraction in transportation geography: its multivariate extension", Canadian Geographer, XII,1.

Whiting, P.D., and Hillier, J.A., 1960, "A method for finding the shortest route through a road network".

Operation Research Quarterly, 11, 1/2.

Wilson, A.G., 1974, Urban and Regional Models in Geography and Planning, John Wiley, London.

Appendix I

Find the Geodesic on a cone given by:

$z(x,y)=1.5\sqrt{x^2+y^2}$ (33)

In spherical coordinates:

$x=r \sin(\theta) \cos(\varnothing)$

$y=r \sin(\theta) \sin \varnothing$

$z=r \cos(\varnothing)$

For a cone $\theta =$ a constant

Let $\sin(\theta)=\alpha$ and $\cos(\theta)=\beta$, then:

$x=\alpha r \cos(\varnothing)$ and $dx=\alpha(\cos(\varnothing) dr - r \sin(\varnothing) d\varnothing)$

$y=\alpha r \sin(\varnothing)$ and $dy=\alpha(\cos(\varnothing) dr - r \sin(\varnothing) d\varnothing)$

$z= \beta r$ and $dz= \beta dr$

If the distance of a geodesic is s; then

$ds^2 = dx^2 + dy^2 + dz^2,$

which implies $ds=(\alpha^2 r^2+(dr/d\varnothing)^2)^{1/2} d\varnothing$(34)

In order to minimize the integral

$s= \int_a^b ds$ (35)

from point a to b, the following Euler equation must be solved:

$dr/d\varnothing=\alpha r((r^2/r_0^2)-1),$ (36)

where r_0 is a constant.

Set $r/r_0 = \sec(u)$ then,

$$dr/d\phi = r_0 \sec(u) \tan(u) (du/d\phi)$$

$$\text{and } \alpha r ((r^2/r_0^2) - 1)^{1/2} = \alpha r \sec(u) \tan(u),$$

which implies $r = r_0 \sec\{\alpha(\phi - \phi_0)\}$ is a solution to the Euler equation (36). If r is substituted into (34) and integrated the following results:

$$ds = (\alpha^2 r^2 + (dr/d\phi)^2)^{1/2} d\phi,$$

$$dr/d\phi = \alpha r \operatorname{cosec}\{\alpha(\phi - \phi_0)\} \cot\{\alpha(\phi - \phi_0)\},$$

$$\begin{aligned} \text{and } ds &= \alpha^2 r_0^2 \{\sec^2(u) + \sec^2(u) \tan^2(u)\}^{1/2} d\phi, \\ &= \alpha r_0 \sec^2(u) d\phi; \end{aligned}$$

$$\begin{aligned} \text{which implies that } s &= \alpha r_0 \int \sec^2(u) d\phi, \\ &= r_0 \tan\{\alpha(\phi - \phi_0)\} \dots\dots\dots (37) \end{aligned}$$

In order to calculate the length of the geodesic of the cone (33) from $P(x,y)=(-1,-1)$ to $(1,1)$ r_0 and ϕ_0 in equation (37) must be found. It is argued that the geodesic from $P(x,y)=(-1,0)$ to $(+1,0)$ has the same length. For this case $r = 1.6414$,
 $\phi = \pi/2$.
The length for this geodesic is $s=3.901$.

University of Alberta Library



0 1620 1713 8379

B30273